Stochastic Control Approach to the Multi-Armed Bandit Problems

Tanut Treetanthiploet

Mansfield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

Acknowledgements

The completion of this thesis would not have been possible without the invaluable assistance given by my supervisor, Sam Cohen. I am deeply indebted to him for his motivational spirit, his mathematical insight, and advice on so many other topics, from coding to the English language and so much more. He has been encouraging, empathetic and inspiring from the first to last day of my study.

I would like to express special thanks to my family especially my mother, Somying Wongsaitong, for her continuous love, support and encouragement from my birth and through the ten years of my studies in the UK. I could not have made it here without her.

I also would like to extend my gratitude to my former teachers, tutors, lecturers and supervisors from school to university for their many years of lessons and advice. Special thanks to C. Mouhot and N. Berestycki for in-depth supervision during my undergraduate study, which allowed me to extend my knowledge to pursue my DPhil; to D. Stewart for being a wonderful Director Of Studies; to M. Ford for his patience in many of my lessons when I first arrived the UK; to S. Sumetkijakan for guiding me through my first mathematical project; and to C. Gunprawit, P. Rungrat, T. Kamtientong and A. Siripin for their generousity and limitless advice in my schooldays.

Thanks also to many of my friends who have supported me through the completion of my DPhil. Special thanks to John (Yufei) for many helpful discussions, to James and Harry for being my cooking brothers and giving unlimited support to help me settle into the British culture, and to Mapraw and Ploy for the valuable friendship and mental support through my study.

Finally, I would like to acknowledge financial support from the Development and Promotion of Science and Technology Talents Project (DPST) of the Government of Thailand.

Abstract

A multi-armed bandit is the simplest problem to study learning under uncertainty when decisions affect information. A standard approach to the multi-armed bandit often gives a heuristic construction of an algorithm and proves its regret bound. Following a constructive approach, it is often possible to find a scenario where following heuristic approaches gives a poor decision.

In this thesis, we consider solving the multi-armed bandit problem from first principles, in terms of stochastic control. We propose two novel approaches to address the multi-armed bandit problem. The first approach is to apply a relaxed control analogy to obtain a semi-closed form approximation to the optimal solution. The proposed model covers a wide range of bandit problems, and the proposed strategy can be computed with a low computational complexity with an empirically strong performance. The second approach focuses on bandits with independent arms and considers the interaction between two aspects of uncertainty: uncertainty aversion and learning. These aspects are in some sense opposite; one is pessimistic, but another is optimistic. To see this interaction, we consider a class of strategies that allows marginal projection on each arm and prove Gittins theorem under nonlinear expectation.

Overall, our proposed approaches provide an understanding of how to make decisions under uncertainty when our decisions determine future information. These results should be helpful as a foundation to combine stochastic control with more modern AI theories.

Contents

List of Figures

Chapter 1

Introduction

Reinforcement Learning is a very active field in Machine Learning and is key to understanding the development of Artificial Intelligence (AI). The theory of Reinforcement Learning is often concerned with solving a Markov decision process using the Bellman equation. The Bellman equation relies on the decision-maker's knowledge of the parameters of the underlying system, which one often needs to estimate in practice. The majority of works on reinforcement learning focus on finding a method to solve or estimate an optimal solution given a probabilistic model (exploitation), whereas learning of the underlying system (exploration) is taken for granted, i.e. the parameter of the underlying system is assumed to be known.

In many real-world situations, we never have perfect knowledge of the underlying system and our decisions could affect the amount of information obtained. An option which provides more information on the underlying system could be sub-optimal according to the estimated parameter, but valuable when making future decisions. The decision-maker now faces a dilemma between choosing the optimal solution evaluated with the estimated parameters (exploitation) and choosing options that reveal more about the parameter, in order to have a more accurate estimate (exploration). This dilemma is known as the exploration vs. exploitation trade-off.

This thesis studies this trade-off in reinforcement learning. To emphasise our primary focus, we will consider a model where we need to choose among finite options where we would know the optimal strategy immediately if we knew what the actual parameter was. In practice, the actual parameter is often unknown, but we need to act to reveal a small piece of information about it. Different actions give different rewards and different information. The mathematical framework for this situation is 'the multi-armed bandit problem'.

1.1 Multi-armed bandit problems

Multi-armed bandit problems (or, in short, bandit problems) are a classic toy example to study decision-making with randomness when the actual distribution is unknown. They were first introduced by Thompson [104] in early 1930s for applications in medical trials (Armitage [6] or Anscombe [5]). They are also useful for experimental design (Berry and Fristedt [15]), along with other areas. A few recent works in finance for portfolio selection can also be found in Huo and Fu [59], or Shen et al. [101].

The basic idea of bandit problems is that we need to decide between K alternatives, which are often referred to as the K arms of a bandit. When an arm is chosen, an observation is generated corresponding to our decision, and we collect the corresponding reward/cost. The objective of the problem is to maximise (or minimise) the total reward (or cost) possibly discounted through time. For clarity, we will consider maximisation in this introductory chapter and we will restrict with this until Chapter 5. From Chapter 6 onward, we will consider the minimisation problem, in order to more simply align with results we use from other fields.

Bandit problems have a long history and have been studied in many disciplines. Most of the research on bandits focuses on proposing a promising algorithm and evaluating its performance rather than defining a problem and solving it. Thus, these algorithms cannot explain at a fundamental level how we should react to information. On the other hand, studying bandits through control theory formulates the problem from first principles and can be used to explain our decision's nature under uncertainty. However, there are a couple of challenges, both computational and theoretical, to this stochastic control approach. We will discuss these challenges and methods to overcome them.

There are two main contributions in this thesis. First, we consider a bandit problem via a relaxed control setting. We use this to establish a near-optimal decision in terms of a semi-index strategy, which can be computed efficiently by solving a fixed point problem. We also test its performance through the simulation in competition with state-of-the-art methods, and see that it does well. The second part of this thesis aims to understand decisions under uncertainty aversion. Due to the time-inconsistency effect, modeling uncertainty aversion is not easy. Thus, we study and propose a relaxation on the concept of optimality. We then prove a version of Gittins' theorem, stating that the index strategy yields an optimal solution under this optimality criteria.

Due to its long history and the different disciplines of researchers studying this problem, the structure of the multi-armed bandit problem is often unclear. Hence, we will start our discussion by giving a summary of the different elements of the problem. We will then discuss related concepts that inspire the contribution of this thesis.

1.1.1 Decisions for the Multi-Armed Bandit Problems

The Multi-Armed Bandit problem is often studied in statistics and computer science communities. Hence, a detailed mathematical formulation of the resulting filtrations and measurability of decisions is often omitted or roughly stated. Essentially, a peculiarity of this problem is that our decisions determine the filtration. Therefore, a carefully stated definition of filtration and the measurability of the decision (or equivalently the control) is required.

Let $(\Omega, \mathbb{P}, \mathcal{F})$ be an underlying probability space equipped with a family of random variables $(Y_t^{(k)})_{k=1,..,K,t\in\mathbb{N}}$ and let $(\zeta_t)_{t\in\mathbb{N}}$ be an independent and identically distributed sequence of $U[0,1]$ random variables. Let $\mathcal{F}_t^U$ be a filtration representing the information available at time t, which will be defined shortly. Finally, let $(R^{(k)})_{k=1,...,K}$ be a collection of reward functions.

In the Multi-Armed Bandit problem, the decision-maker typically needs to make a *random choice*, which is modeled by a process $(A_t)_{t\in\mathbb{N}}$ taking values in a finite set $[K] := \{1, 2, ..., K\}$. The decision-maker does not choose $(A_t)_{t\in\mathbb{N}}$ directly but instead chooses a family of conditional laws $(\Pi_t(\cdot))_{t\in\mathbb{N}}$ of A_t, where $\Pi_t(\cdot) := \mathbb{P}(A_t \in \cdot | \mathcal{F}_{t-1}^U)$.

The process (ζ_t) is considered as a random seed for the strategy Π_t. Each randomised policy Π_t can be represented uniquely by an $\mathcal{F}_{t-1}^U$-measurable random variable U_t taking values in the K-dimensional simplex, $\Delta^K := \left\{u \in [0,1]^K : \sum_{i=1}^K u_i = 1\right\}$.

The connection between Π_t and U_t can be given by $U_t = \big(\Pi_t(\{1\}), ..., \Pi_t(\{K\})\big)$. We will call $(U_t)_{t\in\mathbb{N}}$, a *probabilistic decision*. We may also prescribe the random choice A_t in terms of U_t and ζ_t by

$$A_t = A(U_t, \zeta_t) \qquad \text{where} \qquad A(u, \zeta) := \inf\left\{i : \sum_{k=1}^{i} u_k \geq \zeta\right\}. \tag{1.1.1}$$

At time t, ζ_t is revealed to the decision-maker to determine the action A_t, and we observe the information contained in the random variable $Y_t^{(A_t)}$. The decision-maker then collects an instantaneous reward $R^{(A_t)}(Y^{(A_t)})$.

We define the corresponding filtration for these observations by

$$\mathcal{F}_t^U := \sigma(\zeta_1, Y_1^{(A_1)}, ..., \zeta_t, Y_t^{(A_t)}).$$

Remark 1.1. The observation $(Y_t^{(A_t)})$ and the reward function $(R^{(k)})$ may vary between settings, leading to different 'information structures' in practice. The objective of the multi-armed bandit problem is to maxmise the total collected reward under various criteria. We will discuss this further in later sections.

1.2 Variations of the Multi-armed bandit problem

In the literature, one often refers to the multi-armed bandit problem as a problem where the observation $Y_t^{(k)}$'s are generated from a fixed but unknown distribution parameterised by an unknown parameter Θ. [1] When we choose the kth option at time t, we observe $Y_t^{(k)}$and collect the instantaneous reward $R^{(k)}(Y_t^{(k)})$ which we wish to maximise. The decision-maker needs to infer, on-the-fly, the distribution of the observation $Y_t^{(k)}$ while collecting the corresponding reward given his decision k.

Mathematically, we assume that the (conditional) distribution of the observation $(Y_t^{(k)})$ is given by $(Y_t^{(1)}, ..., Y_t^{(K)})|\Theta, \mathcal{F}_{t-1}^U \sim \mu_\Theta(t)$ which could vary depending on the previous observation. We often assume that the type of the distribution $\mu_\Theta(t)$ is known, but the parameter Θ, describing $\mu_\Theta(t)$, is unknown. The decision-maker needs to use his observations $(Y_s^{(A_s)})_{s=1}^t$ at time t to infer Θ, in order to make a decision and collect the corresponding reward at time $t+1$.

The type of distribution $\mu_\Theta(t)$ (together with the reward function $(R^{(k)})$) distinguishes different types of the multi-armed bandit problem. We will discuss a few classes of the bandit problem related to this thesis.

- **Classical (stationary) bandit**: This is a typical class of bandit that is often considered in the literature. It assumes that there is an underlying parameter $\Theta^{(k)}$ corresponding to each observation $Y^{(k)}$ of the kth arm and μ_Θ does not vary in time.

  In particular, this instance of the bandit problem considers the case when $Y_t \sim_{IID} \mu_\Theta = \bigotimes_{k=1}^K \nu_{\Theta^{(k)}}$. The classical bandit in most of the literature commonly assumes that the reward function $R^{(k)}(y)$ is the identity and the distribution of $\nu_{\Theta^{(k)}}$'s are the same for all k. In this case, we often call the type of the distribution $\nu_{\Theta^{(k)}}$ as the name of bandit problem (e.g., Gaussian bandit, sub-Gaussian bandit, or Bernoulli bandit).

[1] In probability theory, one may interpret bandit as a Markov decision process where the state evolution of each arm is independent. See Section 2.1 to see the connection between these two interpretations.

- **Linear bandit or Correlated bandit**: This model is similar to a classical bandit, but we assume some correlation between the parameters of each arm. The word linear refers to linear dependence in the parameters. Mathematically, we assume that $Y_t \sim_{IID} \mu_\Theta = \bigotimes_{k=1}^{K} \nu_{\Gamma^{(k)}}$ where $\Gamma^{(k)} = (b^{(k)})^\intercal \Theta$ and the $b^{(k)}$'s are known vectors.

- **Adversarial Bandit**: An adversarial bandit is a bandit where we abandon almost all the assumptions on how the rewards are generated. In particular, at each time step, the adversary is free choose $\mu_\Theta(t)$ from any possible measure, and we need to react against it. Since the measure is allowed to change through time without any specific structure, we cannot use our historical observation for statistical inference.

  In the late chapters of this thesis, we will consider an adversarial structure, but this is not the same as the commonly studied adversarial bandit. To be precise, we will allow the adversary to decide $\mu_\Theta(t+1)$ from a set $\mathcal{Q}_t$, where the set $\mathcal{Q}_t$ is given as a part of our problem. In the context of the adversarial bandit, $\mathcal{Q}_t$ is a fixed collection of all possible measures. In our setting, we can model $\mathcal{Q}_t$ through statistical learning and expect it to collapse in time. Thus, we are still in the scope of the learning problem, which is not the case for the adversarial bandit.

- **Contextual Bandit**: It could be the case that the reward or the observation may vary depending on some state of the world, which is known before making the decision.

  In particular, for the contextual bandit, we assume that $Y_t^{(k)} = (C_t, \tilde{Y}_t^{(k)})$ where $\tilde{Y}_{t+1}^{(k)} | \mathcal{F}_t^U \sim \nu_{\Gamma_t^{(k)}}$ and $(C_t)_{t\in\mathbb{N}} \sim \mathcal{C}$ where $\mathcal{C}$ is a known distribution and $\Gamma_t^{(k)} = \psi(k, C_t, \Theta)$ for some known function ψ. In the case when ψ is linear in Θ, we also call this a *linear contextual bandit* or it may simply be referred to as a *linear bandit* in some literature. This is a generalised version of the linear (correlated) bandit described earlier, but in those cases ψ does not depend on C_t.

  Our thesis will restrict the notion of linear bandit to the case where ψ does not depend on C_t. We will also consider the contextual bandit problem in a general setting, but we require the context to satisfy some further assumptions.

  N.B. When the contextual bandit is considered in the literature, they often do not assume the distribution of (C_t) directly. However, they either use the adver-

sarial bandit argument or assume the consistency of their estimated parameter to study the problem and establish a 'regret' guarantee.

1.3 Frequentist vs Bayesian: from Statisticians to Probabilists

In the earlier section, we discussed various instances of bandits where we described the distribution of the observation given a **parameter** Θ by $Y_t|\Theta \sim \mu_\Theta(t)$. In some sense, our previous statement is ambiguous. We state that Θ is a parameter, which is often interpreted as a real number; however, the notion of the conditional law interprets Θ as a random variable. To talk about the interpretation of the parameter Θ, we refer to the two main schools of statistics: frequentist vs. Bayesian.

In the frequentist school, we interpret Θ as an unknown real number. Since probability theory cannot distinguish the word 'known' and 'unknown', Θ shall be treated as a part of the probability measure. Thus, the admissible class of strategies becomes ambiguous, as we want to insist that we cannot explicitly have Θ in our decision. Therefore, we can no longer consider the multi-armed bandit problem as a standard optimisation problem in stochastic control. This is a fundamental reason why one can only propose an algorithm and evaluate a 'regret' bound rather than solving the control problem; in the mainstream (frequentist) literature, such a 'problem' does not exist.

On the other hand, a Bayesian views the parameter Θ as a hidden random variable sampled from a known prior. We then assume the distribution of the observation is simply the conditional law given Θ as described earlier. In this case, the admissible strategy is clear, which is the decision that we discussed in Section 1.1.1. When considering in this framework, we may refer to the word 'known' and 'unknown' parameters to emphasise the nature of the information about the distribution. A 'known' parameter shall be interpreted as a fixed real number equipped to the probability, whereas an 'unknown' parameter shall be interpreted as a hidden random variable sampled from a known prior.

Our thesis studies bandit problems under both statistical inference. We will discuss about the use of these inferences in Section 1.6.

1.4 History and Objectives of the Multi-armed bandit problem

We have already discussed the decision structure and (statistical) interpretation of the parameter for the bandit problem. However, we have not clearly stated the 'problem' yet. In the bandit problem, we wish to choose a strategy to maximise our reward; but in what sense do we maximise? Especially in the frequentist perspective, the maximisation problem is not even well-defined due to the lack of admissible strategies, as discussed earlier.

To give a precise description of what has been done and to refer to other works in this field, we will discuss a few variations on the objective of the bandit problem. To allow the reader to see these variations, we will give a brief history to explain the use of those objectives.

Historically, the multi-armed bandit problem was first considered from a Bayesian perspective by Thompson [104] without explicitly specifying an optimisation problem. Thompson proposes to sample $\hat{\Theta}$ from the posterior of Θ and treat $\hat{\Theta}$ as a fixed parameter to make a decision. This method is later known as 'Thompson Sampling (TS).' [2]

The first mathematical result regarding solving bandit problems was proposed by Gittins and Jones [54, 55] for a classical (stationary) bandit. They also formulate the problem using a Bayesian framework and consider maximising the total discounted reward:

$$V(U,\beta) := \mathbb{E}\Big(\sum_{t=1}^{\infty} \beta^{t-1} R^{(A_t)}(Y_t^{(A_t)})\Big) \tag{1.4.1}$$

where $\beta \in (0,1)$ and $A_t = A(U_t, \zeta_t)$. Maximising (1.4.1) directly requires the use of backward induction with posterior distributions as an underlying state. This method results in the curse of dimensionality (see Section 2.1 for further discussion). In short, as the number of arms increases, the dimension of the parameter $\Theta = (\Theta^{(1)}, ..., \Theta^{(K)})$ increases, and so does the dimension of an underlying state. This makes backward induction computationally intractable.

Gittins and Jones [54] proposed a method to solve (1.4.1) without considering a high-dimensional problem, in the case when (the prior of) $(\Theta^{(k)})_{k=1,...,K}$ are independent. They show that at each time step, one needs to compute the *allocation index* (which is later called *Gittins index*) by considering K optimal stopping problems. Each problem considers the posterior of each arm separately; thus, the dimension

[2] There is also a trivial approach to always choose based on the best estimate, i.e. the 'greedy algorithm.' However, its early role in the history of bandit problems is unclear.

does not increase with the number of arms. The optimal solution to (1.4.1) is then given by always choosing the option which has the maximum index. In particular, Gittins' theorem says that for an independent bandit, we only need to solve K problems with low dimension rather than solving one problem with high dimension.

Independent of Gittins, Lai and Robbins [70] considered the multi-armed bandit problem from a frequentist perspective. They define the regret of the strategy U given a parameter Θ by

$$\mathcal{R}(U,T,\Theta) := \mathbb{E}\Big(\sum_{t=1}^{T}\big(R^{(k^*)}(Y_t^{(k^*)}) - R^{(A_t)}(Y_t^{(A_t)})\big)\Big|\Theta\Big) \tag{1.4.2}$$

where $A_t = A(U_t,\zeta_t)$ and $k^* = \arg\max_k \mathbb{E}\big(R^{(k)}(Y_t^{(k)})\big|\Theta\big)$. [3]

They showed that, for the classical bandit with $Y_t^{(k)} \sim_{IID} \nu_{\Theta^{(k)}}$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathcal{R}(A,T,\Theta)|\Theta]}{\log(T)} \geq \sum_{k=1,k\neq k^*}^{K} \frac{\mathbb{E}\Big(\sum_{t=1}^{T}\big(R^{(k^*)}(Y_t^{(k^*)}) - R^{(k)}(Y_t^{(k)})\big)\Big|\Theta\Big)}{I(\Theta^{(k)},\Theta^{(k^*)})} \tag{1.4.3}$$

where $I(\theta,\theta')$ is the Kullback–Leibler distrance between the distributions ν_θ and $\nu_{\theta'}$.

Even if Gittins' theorem gives a dimensional reduction for maximising (1.4.1), computing Gittins index is not an easy task. Nonetheless, an index strategy is still a simple and reasonable analogy to implement a decision. Agrawal [1] used Gittins index theorem to inspire a method to tackle the classical bandit by defining an index strategy via the Upper Confidence Bound with appropriate scaling. This method is later known as the *UCB algorithm*. He then shows that this strategy yields an asymptotic regret of order $\mathcal{O}(\log T)$, i.e. he shows that his strategy obtains the reverse of the inequality (1.4.3) up to a multiplying factor.

Due to the difficulty in the analysis of Gittins' theorem and its computation, most practical works on bandits follow a similar approach to Agrawal [1]. They often consider a specific bandit problem, establish the lower bound as in Robbins and Lai [70], and propose an algorithm that achieves or nearly achieves such bound.

The consideration of the objective extends to various formulations. One may consider the term inside the expectation (1.4.2) and give a bound in the high-probability sense or one may replace $\big(R^{(k^*)}(Y_t^{(k^*)}) - R^{(k)}(Y_t^{(k)})\big)$ by $\mathbb{I}\big(R^{(k^*)}(Y_t^{(k^*)}) = R^{(k)}(Y_t^{(k)})\big)$ (see e.g. Burtini [22] for the discussion). Another common variation is to consider the Bayesian regret;

$$\mathcal{BR}(U,T,\pi) := \mathbb{E}_\pi[\mathcal{R}(U,T,\Theta)] \tag{1.4.4}$$

[3] Here, we restrict our notation to a Bayesian perspective by denoting Θ as a random variable, to give consistent notation. One can see (1.4.2) as a classical expectation given a parameter $\Theta = \theta$ through the Doob–Dynkin lemma.

where we denote π to emphasise the prior of Θ.

Apart from Gittins' result (and the follow-up literature on its generalisation) on the optimal solution, optimal solutions to other bandit variations, where the arms are not independent, have not yet been studied in detail. Therefore, optimality becomes the main focus of our thesis. We will establish results regarding an optimal solution to (1.4.1).

1.5 Uncertainty

In earlier sections, we discussed how the multi-armed bandit problem is set up and its use of observations to infer the true paramater Θ. The inference for the parameter Θ comes with statistical *uncertainty* in our estimate.

In the classical bandit problem, we only observe the information of $\Theta^{(k)}$ when the kth arm is played. Hence, if the kth arm was played too seldom, we may have an incentive to play the kth arm more, in order to learn its parameter. In this way, we view uncertainty positively, i.e., we are optimistic toward it. In contrast, the stochastic control community draws on economic considerations, which generally suggest a bias against uncertainty.

In this section, we will discuss these different aspects of uncertainty that we face in our decision making.

1.5.1 Uncertainty in the Multi-armed Bandit

When considering a classical bandit, it might be better to play a choice where we know less, in order to gain more information. Gittins' theorem shows that the optimal decision for the classical bandit can be made by choosing the maximal index, say α. It was shown, in various works, based on different assumptions (e.g. Gittins [56], Brezzi and Lai [21], Russo [95]), that the index α can be written in the form

$$\alpha = \text{Estimated mean of the reward} + \text{Learning Reward}. \tag{1.5.1}$$

where the 'Learning Reward' is positive and increases with the posterior variance of the estimate. The posterior variance can be interpreted as the 'statistical uncertainty' of our estimate. In particular, (1.5.1) means that we are optimistic toward uncertainty.

The principle of putting uncertainty as an additional learning reward and considering an index strategy of the form (1.5.1) is a standard and popular approach to

tackling bandit problems in the literature. This approach is commonly known as the *UCB algorithm* (see section 2.2.3).

Similar to the UCB approach, the Knowledge Gradient (KG) [100] (see section 2.2.4) is another approach which can be used to tackling bandit problems. We see, in [100, Equation 13] that the resulting strategy for the Gaussian bandit also results in an index strategy as described by (1.5.1).

We will also see in Chapter 2 that many other approaches for bandits also favour uncertainty in the sense that additional randomness in the decision is preferred. In particular, they prefer U_t not yield deterministic decisions. (This is why we allow an additional process (ζ_t) in Section 1.1.1 to reflect this preference.)

Overall, we may argue from the above observations on the theory on bandits that they often display a bias in favour of uncertainty. [4]

1.5.2 Ellsberg paradox and Uncertainty aversion

We have discussed the positive aspect of uncertainty in bandits. We now discuss uncertainty aversion, a more familiar concept in the stochastic control community.

Let consider the following scenario observed by Ellsberg [44]. Suppose a box contains 100 red balls and 200 other balls that are either black or yellow. It is unknown how many black or how many yellow balls there are, but the total number of black balls plus the total number of yellow is known to be 200. Suppose that a single ball is drawn from a box, and we need to choose one of the two options to bet.

Which of these options shall we choose?

A: Win £100 if a red ball is drawn
B: Win £100 if a yellow ball is drawn

Also, which of these options shall we choose?

C: Win £100 if a red or black ball is drawn
D: Win £100 if a yellow or black ball is drawn

A survey shows that most people strictly prefer A to B but D to C, which violates classical probability theory.

Let r, y and b be the probability of drawing red, yellow and black balls respectively. By the utility theory (von Neumann and Morgenstern [106]), one can see that strictly preferring A to B implies that

$$rU(100) + (1-r)U(0) > yU(100) + (1-y)U(0). \tag{1.5.2}$$

[4] We will discuss further in Chapter 3 that bandits do not generally like uncertainty, but it is the value of information that they care about, and uncertainty affects the information value.

On the other hand, strictly preferring D to C implies that

$$(r+b)U(100) + (1-(r+b))U(0) > (y+b)U(100) + (1-(y+b))U(0). \tag{1.5.3}$$

By rearranging (1.5.3), one can see it yields the reverse inequality to (1.5.2).

This 'irrational' evaluation in decisions can be understood via our perception of the information of the balls. When making decisions, people generally have a strict preference for options that they know well. Since the classical work of Knight [68] and Keynes [65], there has been a stream of thought within economics and statistics that focuses on the difference between the randomness of an outcome and lack of knowledge of its probability distribution (sometimes called 'Knightian uncertainty'). This lack of knowledge is often related to estimation, as the probabilities used are often based on past observations. In the above example, we do not know the exact number of yellow/black balls.

This lack of knowledge is often considered in the modern stochastic control literature. We often formulate the problem assuming more than one probability measure which describes the underlying. The (robust) optimal solution is defined as the strategy that optimises under the worst-case scenario. This agrees with the example above where we prefer A to B but D to C. The worst-case (possible) probability that B (or C) occurs is lower than the probability that A (or D) occurs, which is known to be 1/3 (or 2/3).

We can see that this interprets uncertainty in a negative way, unlike the interpretation common for bandits. For simplicity, we will not consider this type of uncertainty until Section 6, where we will mix these two concepts of uncertainty through the bandit problem.

1.6 Approaches for controls determining filtrations

We have discussed in Section 1.1.1 that our decisions determine our filtration. This phenomenon results in the difficulty of analysing the multi-armed bandit in general. The two approaches of this thesis handle the filtration effect differently, allowing different statistical inference and bandit structures to be analysed. This section will motivate these approaches and specified the related bandit types and their statistical inference.

The first approach (Chapters 3, 4 and 5) considers estimation in a Bayesian framework and handles the filtration effect through a Markov decision process.

Suppose we have an unknown parameter Θ such that

$$\mathcal{L}(Y_{t+1}, Y_{t+2}, ...|\Theta) = \mathcal{L}(Y_{t+1}, Y_{t+2}, ...|\Theta, Y_1, ..., Y_t) \qquad \text{for all } t \geq 0, \tag{1.6.1}$$

where $\mathcal{L}$ is the likelihood function.

Let X_t^U be a posterior parameter of Θ at time t. It follows that X_t^U is a (Bayesian) sufficient statistic for Θ given $\mathcal{F}_t^U$, i.e. $\mathcal{L}(\Theta|\mathcal{F}_t^U) = \mathcal{L}(\Theta|X_t^U)$.

Hence, it follows from (1.6.1) that $\mathcal{L}(Y_{t+1}, Y_{t+2}, ...|\mathcal{F}_t^U) = \mathcal{L}(Y_{t+1}, Y_{t+2}, ...|X_t^U)$. By Bayes' rule, we can show that (X_t^U) forms a Markov process with respect to $(\mathcal{F}_t^U)$.

Using this approach, we can ignore the filtration effect and see the bandit problem as a Markov decision process, given that (1.6.1) is satisfied. In particular, this approach can handle the classical bandit and the correlated bandit described in Section 1.2. Furthermore, it could be easily extended to study contextual bandits given that the context (C_t) is a Markov process.

The second approach (Chapters 6, 7, 8 and 9) overcomes the filtration effect by establishing Gittins' theorem, which states that when the observations $(Y_t^{(k)})_{t\in\mathbb{N}}$ are independent for each k, the optimal solution to (1.4.1) can be given by an index strategy where the index can be computed separately for each k. In particular, we only need to analyse the filtration corresponding to the information from each arm, $\mathcal{F}_s^{(k)} := \sigma(Y_1^{(k)}, ..., Y_s^{(k)})$ which is not affected by our control, due to the independence assumption.

The challenge of the second approach is to extend the classical Gittins' theorem to a robust version defined under nonlinear expectation. This extension requires careful analysis of the probability space, filtration, and control strategy to ensure measurability and consistency of the problem.

This robust version allows us to study bandits with an adversary as discussed in Section 1.2. At each time step, the adversary chooses $\mu_\Theta(t+1)$ from a set $\mathcal{Q}_t$. Under this structure, we can encode either Bayesian or frequentist statistics to describe the set $\mathcal{Q}_t$. In particular, if we are a Bayesian, we can construct $\mathcal{Q}_t$ corresponding to the posterior distributions obtained from multiple priors. If we are a frequentist, we can construct $\mathcal{Q}_t$ corresponding to the confidence interval for the parameter.

1.7 Outline of the thesis

This thesis proceeds by briefly reviewing the relevant Gittins' theorem and the algorithms for bandits in Chapter 2. We then discuss our contribution to seeing bandits as a stochastic control problem. As discussed above, we provide two main approaches,

corresponding to computational and theoretical perspectives. The first approach is computationally friendly and is useful in practice. We will see the bandit as a relaxed control problem and derive an approximate optimal decision. The second approach concerns a more delicate theoretical perspective. This approach allows us to understand the interaction between learning and uncertainty aversion in decision-making. We will outline the structure of this thesis based on these two perspectives.

Asymptotic Random Control (ARC) algorithm

The first half of this thesis results in the Asymptotic Random Control (ARC) algorithm. We give a discussion of this approach in Chapters 3, 4 and 5 where the presented material is based on two papers: 'Asymptotic Randomised Control with applications to bandits' [34] and 'Correlated Bandits for Dynamic Pricing via the ARC algorithm' [36] written by the author and S.N. Cohen.

The initial idea of this approach is discussed in Chapter 3, where we describe the posterior distribution of Θ parameterised by a pair (m, d) and see it as a Markov underlying process, as discussed in Section 1.6.

By considering the relaxed control problem and exploring the structure of the information flow, we can give a semi-closed form approximation to the optimal control problem with an objective corresponding to (1.4.1). The resulting approximate strategy has a natural interpretation where we quantify the 'value of instantaneous reward' and the 'value of information.'

After finishing the theoretical discussion in Chapter 3, we consider a numerical simulation in Chapter 4 and find that our approximate control performs competitively with other approaches for the bandit problem.

The requirement for the existence of the posterior parameter (m, d) fundamentally restricts the ARC algorithm to the case where the observations and Θ are a conjugate pair. In Chapter 5, we extend the application of the ARC algorithm to the case where observations come in batches from an exponential family without requiring a conjugacy assumption. The parameter here describes the observations through a generalised linear model. We apply this model to the dynamic pricing problem, and find empirically that the ARC algorithm performs relatively well compared with other methods.

Robust Gittins' Theorem

The second half of this thesis proves Gittins' theorem under uncertainty aversion. This result is discussed in Chapter 6, 7 and 8 based on the paper 'Gittins' theorem under

uncertainty' written by the author and S.N. Cohen [35]. We also provide a further approximate asymptotic analysis of this result (in continuous time) in Chapter 9.

The key idea of this result is to answer the question 'how do the two aspects of uncertainty described in Section 1.5 interact with each other?'.

The typical approach to model uncertainty in modern stochastic control is through a 'nonlinear expectation.' However, due to the difficulty that the decision determines the filtration, the nonlinear expectation faces some form of inconsistency. In Chapter 6, we describe this difficulty and propose relaxation of the optimality requirement to overcome it.

In Chapter 7, we give an outline of the proof of a robust (or uncertainty aversion) version of Gittins' theorem. We also provide a numerical computation of the robust Gittins index which illustrates an interaction between uncertainty aversion and learning benefit, connecting to behavioural economics. The full proof of this theorem is provided in Chapter 8.

Finally, in Chapter 9, we consider a continuous-time version of Gittins index, review the existing results, and (semi-formally) extend some of the results to obtain the asymptotic behaviour of a robust Gittins index in the Gaussian case. This explains some observations that we see in the numerical evaluation of the robust indices in Chapter 7 which suggests that uncertainty aversion will eventually dominate learning.

Chapter 2

Overview of the Multi-armed bandit problem

Before discussing our approaches, we will overview what has been done for the multi-armed bandit problem.

In the first half of this chapter, we discuss the multi-armed bandit problem from the perspective of Gittins index theorem. We then review related works from both theoretical and computation aspects of the index. The second half of this chapter describes approaches to the multi-armed bandit problem from the statistics and computer science disciplines. We give a brief introduction to a few commonly used approaches for the bandit problem in practice. We will later compare these approaches with the algorithm constructed based on our new analogy.

2.1 Gittins index theorem

As briefly mentioned in Section 1.4, solving the bandit problem using the Bellman equation is usually computationally intractable. To overcome this difficulty, Gittins and Jones [54] proposed the first mathematical result to solve this problem. We shall explore Gittins and Jones' result in greater detail. For simplicity, we will discuss the theorem in the same setting that it was originally stated and later connect it to our stated framework.

Suppose that there are K options to choose, based on the state $(X_t^{(1)}, ..., X_t^{(K)})$ taking values in χ^K. When the kth option is chosen at time t, we receive a reward $r^{(k)}(X_t^{(k)})$ and only state $X_t^{(k)}$ may evolve. Other states remain the same. Suppose that we would like to maximise the total discounted reward

$$v(x) := \sup_A \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t r^{(A_{t+1})}(X_t^{(A_{t+1})}) \Big| X_0 = x\Big], \tag{2.1.1}$$

where $\beta \in (0,1)$ and (A_t) is our control process taking values in $\{1,...,K\}$[1]. It follows from the Dynamic Programming Principle that

$$v(x) = \max_{k=1,...,K} \Big\{ r^{(k)}(x^{(k)}) + \beta \mathbb{E}\Big[v(X_1) \Big| X_0 = x, A_1 = k \Big] \Big\}. \tag{2.1.2}$$

We can see that the Bellman equation (2.1.2) above is (at least) k-dimensional and thus computationally intractable.

However, Gittins and Jones [54] exploited the independence structure and showed that for the problem described above, solving (2.1.2) directly is not required.

Theorem 2.1 (Gittins index theorem)**.** *Let (X_t) be a process with an independent evolution described earlier. An optimal control for* (2.1.1) *can be given by a feedback control $A(x) = \arg\max_{k=1,...,K} \gamma^{(k)}\big(x^{(k)}\big)$ where*

$$\gamma^{(k)}\big(x^{(k)}\big) := \sup_{\tau>0} \frac{\mathbb{E}\Big[\sum_{t=0}^{\tau-1} \beta^t r^{(k)}\big(X_t^{(k)}\big) \Big| X_0^{(k)} = x^{(k)} \Big]}{\mathbb{E}\Big[\sum_{t=0}^{\tau-1} \beta^t \Big| X_0^{(k)} = x^{(k)} \Big]}. \tag{2.1.3}$$

Equivalently, Weber [108] also showed that

$$\gamma^{(k)}\big(x^{(k)}\big) = \sup \Big\{ \gamma : \sup_{\tau>0} \mathbb{E}\Big[\sum_{t=0}^{\tau-1} \beta^t \big(r^{(k)}(X_t^{(k)}) - \gamma \big) \Big| X_0^{(k)} = x^{(k)} \Big] \geq 0 \Big\}. \tag{2.1.4}$$

Here, the supremum is taken over positive stopping times τ. The quantity $\gamma^{(k)}\big(x^{(k)}\big)$ is called the Gittins index.

Remark 2.1. The setting for Gittins index theorem is in some sense more general than a classical multi-armed bandit problem. In the classical bandit problem, the observation $(Y_t^{(1)},...,Y_t^{(K)})$ is generated from a fixed distribution $\bigotimes_{k=1}^K \nu_{\Theta^{(k)}}$ where Θ is unknown. Given that the $\Theta^{(k)}$'s are independent (under our considered prior), we can see $X_t^{(k)}$ as the posterior distribution of $\Theta^{(k)}$ at time t. For example, suppose that our belief at time t is $\Theta^{(k)} \sim N(M_t^{(k)}, D_t^{(k)})$ and $(\Theta^{(k)})_{k=1}^K$ are independent (under our considered prior). When the kth arm is chosen at time $t+1$, we observe $Y_{t+1}^{(k)} \sim N(\Theta^{(k)}, \sigma_k^2)$. After the play, $(M_t^{(k)}, D_t^{(k)})$ evolves;

$$M_{t+1}^{(k)} = \frac{\big(D_t^{(k)}\big)^{-1} M_t^{(k)} + \sigma_k^{-2} Y_{t+1}^{(k)}}{\big(D_t^{(k)}\big)^{-1} + \sigma_k^{-2}} \qquad \text{and} \qquad D_{t+1}^{(k)} = \frac{1}{\big(D_t^{(k)}\big)^{-1} + \sigma_k^{-2}}.$$

[1]We can still allow (A_t) to be random through (U_t) as in Section 1.1.1. However, this results in a bang-bang control problem which yields the same Bellman equation as (2.1.2).

The parameter $X_t^{(k)} = (M_t^{(k)}, D_t^{(k)})$ can be seen as the underlying state and the reward $r^{(k)}(m^{(k)}, d^{(k)}) := \mathbb{E}\big[R^{(k)}(Y^{(k)})\big|\Theta^{(k)} \sim N(m^{(k)}, d^{(k)})\big]$ where $Y^{(k)} \sim N(\Theta^{(k)}, \sigma_k^2)$ and $R^{(k)}$ is some reward function.

The underlying state $(X_t^{(k)})$ allows Gittins' theorem to deal with a bandit problem even if we cannot describe the bandit through a parameter Θ. In particular, we can use Gittins' theorem to study purely probabilistic problems without parameter learning, for example, the scheduling problem. Nonetheless, the independence requirement does not allow Gittins' theorem to tackle a bandit problem when there is some correlation in the observations between different arms.

Remark 2.2. It is worth pointing out that many classical works on the Gittins index consider the expected reward $r^{(k)}(X_t^{(k)})$ as an instantaneous reward from our decision at time t. In particular, we know exactly the reward we get when making the decision. In contrast, in the more practical bandit literature, we often use $R^{(k)}(Y_{t+1}^{(k)})$ as the reward from the decision made at time t, i.e. the reward is random. These modeling choices can be seen as being equivalent through the tower property in the classical linear expectation with $\mathbb{E}[R^{(k)}(Y_t^{(k)})|X_t^{(k)} = x] = r^{(k)}(x)$. We emphasise this difference here, as it could lead to confusion later in our discussion.

In Chapters 6, 7 and 8, we will restrict to the 'adapted' observation (Y_t) rather than the 'predictable' observation (X_t). This is because the tower property cannot be assumed for an 'inconsistent nonlinear expectation'. We will also find that this difference leads to slightly different summing indices when considering this index under 'nonlinear expectation' in these later chapters.

2.1.1 Proofs of Gittins index theorem

The original proof of Gittins' theorem [54] involves a complicated interchange of arguments. There are several follow-up papers that give alternative proofs of Gittins' Theorem based on different interpretations (see Whittle [110], Weber [108] and El Karoui and Karatzas [63]). A review of these proofs can be found in Frostig and Weiss [52].

In El Karoui and Karatzas [63], they consider the proof of Whittle [110] together with the dynamic allocation strategy of Mandelbaum [77]. They prove a generalised version of Gittins' theorem without a Markovian assumption. El Karoui and Karatzas later extend the result to continuous time [42] using Mandelbaum's argument in continuous-time [78]. This generalised the earlier results of Karatzas [62], where Gittins index is defined in continuous time for a diffusion process. Bank and El

Karoui [11] developed a stochastic representation to represent a given process using a prescribed function and a running max process which is closely related to Gittins index (in continuous-time). Bank and Küchler [13] later use this to establish a simple proof for Gittins' theorem in continuous-time.

2.1.2 Computation of Gittins index

We have already discussed the solution to (2.1.2) in the form of a Gittins index (2.1.3) or (2.1.4). However, computing the index is also not easy. So far, there is no general closed-form solution to these equations for the discrete time model. Nonetheless, a few approaches can be used to estimate the index. We will give a brief discussion here; the reader can refer to the cited works for further detail.

Since Gittins index only depends on a single-arm process, we will omit a superscript (k) for notational simplicity.

Computation of Gittins index for finite state process: In the case when the state χ is finite, Chakravorty and Mahajan [24] gives a review of the approaches to compute the index through (2.1.3) by explicitly considering the preference on each state of χ and constructing an explicit optimal stopping time. The computation complexity increases with the size of χ and cannot be extended to study our described bandit problem for learning as the state space is generally infinite.

Approximation of Gittins index with Gaussian reward: Gittins [56] proposed a method to estimate Gittins index when the reward is Gaussian with unknown mean[2] as discussed in Remark 2.1 with $R(y) = y$. By scaling and translating normal distribution appropriately, Gittins obtains a decomposition of the index

$$\gamma(m, d, \sigma, \beta) = m + \sigma\gamma(0, d, 1, \beta)$$

where $\gamma(m, d, \sigma, \beta)$ is the Gittins index with (m, d) as the state of the posterior distribution, σ as the standard deviation of the observation and β as the discount factor of the problem. It is then proposed in Gittins [56] to consider using standard backward induction to compute the two dimensional function $\gamma(0, d, 1, \beta)$.

Recently, Russo [95] also applied a similar technique and showed that that

$$\gamma(m, d, \sigma, \beta) = m + \sqrt{d}\Phi^{-1}(\beta) + o(1) \qquad \text{as} \qquad \beta \to 1.$$

[2] In fact, he also explains the estimation for other simple cases.

In particular, Gittins index is asymptotically (as $\beta \to 1$) the Bayesian upper bound of the credible set of level β. This gives us an insight connections between Gittins index theorem and the Bayes UCB-algorithm which will be discussed in Section 2.2.3. The result of Russo verifies an approximation result of Brezzi and Lai [21] which uses a continuous-time argument to approximate the index (see Chapter 9 for further discussion).

2.1.3 Restless bandit and Whittle index

The key assumption for Gittins' theorem is that only the state of the played arm may evolve. Other states of the bandit must remain the same. Whittle proved Gittins index theorem using a simple retirement option [110]. In addition to his proof, Whittle [111] also proposed a more general version of the Gittins index, which is later called the 'Whittle index.'

In his model, he assumes that, when the kth option is chosen, we receive an (active) reward $r_a^{(k)}(X_t^{(k)})$ and the state $(X_t^{(k)})$ evolves with the transition $P_a^{(k)}$. On the other hand, if the kth option is not chosen, the kth arm yields a (passive) reward $r_p^{(k)}(X_t^{(k)})$ and the state $(X_t^{(k)})$ evolves according to $P_p^{(k)}$.

The Whittle index of the kth arm is defined to be the compensating reward needed to leave the kth arm passive. In particular, let us consider the situation when there are two arms and we need to choose one of these arm to play at each time step. The first arm yields a reward in the same manner as the kth arm described above. The second arm gives a constant reward γ only when it is played. It follows from the dynamic programming principle that the corresponding value function is given by

$$V^{(k)}(x^{(k)}) = \max\Big[r_a^{(k)}(x^{(k)}) + \beta P_a^{(k)} V^{(k)}(x^{(k)}), \big(r_p^{(k)}(x^{(k)}) + \gamma\big) + \beta P_p^{(k)} V^{(k)}(x^{(k)})\Big]. \tag{2.1.5}$$

For each $\gamma \in \mathbb{R}$, we can define a set $D^{(k)}(\gamma) \subseteq \chi$ for the states where it is optimal to leave the first arm passive, i.e. the state $x^{(k)}$ such that the maximum of (2.1.5) is achieved by the right hand side.

Definition 2.1. We say that the kth arm is *Whittle indexable* if the set $D^{(k)}(\gamma)$ increases monotonically from ϕ to χ as γ increases from $-\infty$ to ∞. If the arm is Whittle indexable, the *Whittle index* $\gamma_W^{(k)}(x)$ is defined to be the value of γ such that both quantities in the maximum (2.1.5) are equal.

Remark 2.3. In the case where $P_p^{(k)}$ is the identity (i.e. the state of the unplayed arm does not change), the arm is always indexable. In the special case where $r_p^{(k)}(x^{(k)}) = 0$, we reduce to the case where Whittle index is equivalent to Gittins index.

In general, it is not the case that all bandits are Whittle indexable (see Whittle [111] for an example). Some research on bandits concerns proving indexability rather than proving optimality (see e.g., Caro and Gupta [23], Fryer and Harms [53]) and proposed the Whittle index strategy as a reasonable strategy for the decision. This is because Whittle index has a simple interpretation as the average benefit from an arm. Unfortunately, this strategy is not necessarily optimal (see Weber and Weiss [109]); however, Weber and Weiss show that it is asymptotically optimal when the number of arms converges to infinity.

2.2 Algorithms for the Multi-armed bandit

In the earlier section, we discussed Gittins index theorem. This result overcomes the difficulty due to the curse of dimensionality for solving the full Bellman equation (2.1.2). However, computing Gittins index is not an easy task as it involves a non-standard optimal stopping problem. Moreover, Gittins index (or even Whittle index) relies heavily on the assumption that the evolutions of each arm are independent, i.e. the filtrations corresponding to observations restricted to each arm (given a probability measure) are independent.[3] The independence requirement restricts Gittins' result to have a limited use in the real world applications.

Over decades of research on bandits, many approaches can be used to tackle the multi-armed bandit problem. The analysis of these approaches often focus on establishing the order of the regret under a specific setting as discussed in Section 1.4. [4]

In this section, we presents the key idea of some selected approaches and how one can apply them in practice. The reader should refer to the cited works, the survey of Burtini et al. [22] or Lattimore and Szepesvári's book [72] for further analytical detail and the modification of these approaches, including other alternative methods.

2.2.1 Greedy algorithm

The greedy algorithm is the simplest heuristic method to solve a general machine learning or reinforcement learning problem (Sutton and Barto [103]). The Greedy

[3] Equivalently, in a Markovian framework, the 'independence' means that when the kth arm is played, only state $X_t^{(k)}$ of X_t may evolve with the transition that does not depend on other states $(X^{(i)})_{i\neq k}$. Moreover, our instance reward only depends on $X_t^{(k)}$.

[4] The regret analysis of Gittins index is also considered in Lattimore [71] where he varies its definition to consider finite horizon problems.

algorithm's central idea is to always choose the option that maximises the expected (instantaneous) reward.

As this strategy may not allow the decision-maker to learn about the system much, as an alternative, one may allow an ϵ probability to choose an action uniformly at random, i.e., we consider

$$A_t^{Greedy} := \begin{cases} \arg\max_{k=1,\dots,K} \mathbb{E}\Big[R^{(k)}(Y_{t+1}^{(k)}) \Big| \mathcal{F}_t^U \Big] & ; \quad \text{with probability } 1-\epsilon \\ U_t \sim \text{Uniform}(1,\dots,K) & ; \quad \text{with probability } \epsilon \end{cases}.$$

This variation of the greedy algorithm is often known as ϵ-greedy algorithm.

For the classical bandit problem, Even-Dar et al. [45] give an analysis on the upper bound of the number of times that we play suboptimal arm in the high probability sense. A variation in Vermorel and Mohri [105] also allows ϵ to decay in time.

In a more general framework of bandits, Rusmevichientong et al. [93] studied the regret of the pure greedy algorithm for a one-dimensional correlated bandit and proved its regret. Rusmevichientong and Tsitsiklis [94] later extend this bandit structure to a higher dimension where the proposed greedy-based algorithm requires some 'force exploration' in a cycle to guarantee the regret bound.

2.2.2 Force Exploration

Force exploration is another classical approaches for machine learning. In this approach, the decision-maker forces himself to explore for some fixed number of trials, instead of pure randomness as in the ϵ-greedy algorithm. The name of this strategy varies in the literature. Fundamentally, the decisions are split into two phases: the first phase is for exploration and the second phase is for exploitation.

In the exploration phase, the decision-maker may either choose uniformly at random or make a deterministic decision to try all options with an equal number of trials. In the exploitation phase, the decision-maker chooses the option that is expected to be the best, as in the greedy algorithm.

These two phases are commonly split by choosing a parameter $\epsilon \in (0,1)$. The first $\lfloor \epsilon T \rfloor$ trials are considered as an exploration phase where T is the total number of trials. The rest is considered as an exploitation phase. The regret analysis of this approach for a classical sub-Gaussian bandit can be found in [72, Chapter 6] where they call it the Explore-Then-Commit (ETC) algorithm. We will consider this algorithm in our simulation.

In some literature, the exploration phase may be distributed among the trial; for example, Robbins [91], and Rusmevichientong and Tsitsiklis [94]. One may also mix

the force exploration with another algorithm to guarantee sufficient exploration to ensure that the analytical result holds (see, e.g., Cowan et al. [38]).

2.2.3 Upper Confidence Bound (UCB) algorithm

The Upper Confidence Bound (UCB) algorithm is the first non-trivial approach with a low computational cost to tackle the bandit problem. As discussed in Section 1.4, the UCB algorithm was introduced by Agrawal [1] inspired by Gittins' index [54, 55] to achieve the asymptotic regret bound of Lai and Robbins [25]. The connection between the UCB algorithm and Gittins index was later established in Brezzi and Lai [21] and Russo [95]. There are various versions of the UCB algorithm. The version commonly considered in the literature is proposed in Auer et al. [10] for the classical bandit problem where the reward and the observation are the same. [5] The decision for the commonly used UCB algorithm can be described by

$$A_t^{UCB} := \arg\max_k \left(\hat{r}^{(k)}(t-1) + \lambda \sqrt{\frac{\log t}{n^{(k)}(t-1)}} \right)$$

where $\hat{r}^{(k)}(t-1)$ is the average reward and $n^{(k)}(t-1)$ is the number of times that the kth arm has been played. The parameter λ can be varied by the user. The theoretical regret guarantee for sub-Gaussian bandit holds for $\lambda = 2\sigma$, where σ is a variance proxy of $\hat{r}^{(k)}(t-1)$ for all k. Nonetheless, empirical results suggest the use of a lower value of λ.

The main idea of the UCB algorithm is to make the decision based on the index α, where α can be decomposed into two parts:

$$\alpha = \text{Exploitation gain} + \text{Learning premium}. \tag{2.2.1}$$

The exploitation gain is often given in terms of the expected (instantaneous) reward of the chosen choice and the learning premium corresponding to the uncertainty (precision) of the estimated reward with an appropriate rescaling.

The learning premium varies among different settings and the type of regret they wish to establish. One may allow the learning term to capture distribution structures, as in Maillard et al. [76] or Cowan et al. [38]. One may also consider different learning terms to achieve a min-max bound, as in Audibert and Bubeck [9].

In a more general setting, Filipi et al. [47] extends the UCB algorithm to consider a correlated bandit where the reward is generated from a generalised linear model.

[5] Auer et al. [10] also proposed an improved version called 'UCB-tuned' which is also commonly used in practice.

Li et al. [75] also use the UCB algorithm to tackle a contextual bandit problem with linear structure.

The examples considered above consider the UCB algorithm by interpreting bandits from the frequentist perspective. Kaufmann et al. [64] proposed the UCB algorithm in a Bayesian framework. This algorithm is called the *Bayes-UCB algorithm*, where we choose based on the upper-bound of the credible set, i.e.

$$A_t^{Bayes-UCB} = \arg\max_i Q\Big(p, R^{(k)}(Y_t^{(k)})\Big|\mathcal{F}_{t-1}^U\Big),$$

where $Q(p, X)$ is the p-quantile of X.

Kaufmann et al. [64] prove a theoretical (frequentist regret) guarantee for the classical (independent) Bernoulli bandit when $p = 1 - 1/\big(t(\log T)^c\big)$ and $c \geq 5$ where T is the horizon; their simulations suggest that $c = 0$ performs well in practice.

Even if theoretical regret analysis of the Bayes-UCB restricts to a simple case, the extension of this algorithm to general setting is natural. Other UCB type algorithm does not have a natural extension. Therefore, in most of our simulations, we will use the Bayes-UCB algorithm as a candidate.

2.2.4 Knowledge Gradient (KG) method

Ryzhov, Powell and Frazier [100] proposed an alternative method to solve the multi-armed bandit problem. They adopt a similar setting to Gittins [54, 55, 56], but their analogy do not strictly require a separate parameter for each arm.

In particular, they consider (X_t) as a state representing posterior parameters, as in Section 2.1, and try to maximise the total discounted reward

$$v(x) := \sup_A \mathbb{E}\Big[\sum_{t=1}^{\infty} \beta^{t-1} R^{(A_t)}(Y_t^{(A_t)})\Big|X_0 = x\Big] = \sup_A \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t r^{(A_{t+1})}(X_t^A)\Big|X_0 = x\Big],$$

where $r^{(k)}(x) := \mathbb{E}[R^{(k)}(Y_1^{(k)})|X_0 = x]$ and (X_t^A) evolves according to the control A.

As discussed in Section 2.1, solving for $v(x)$ is computationally intractable. Hence, they consider a slight modification of the problem using one-step look ahead strategy. In particular, they assume that we only learn after the current step and then choose the same decision afterward. By considering the value function corresponding to this modification, they obtain an index type strategy where the index is given by

$$\gamma_{KG}^{(k)}(x) = r^{(k)}(x) + \Big(\frac{\beta}{1-\beta}\Big)\mathbb{E}\Big[\max_{k=1,\dots,K} r^{(k)}(X_1)\Big|X_0 = x, A_1 = k\Big]. \tag{2.2.2}$$

The strategy of the Knowledge Gradient algorithm is to play the arm which maximise $\gamma_{KG}^{(k)}(x)$.

In [100], the setting is only restricted to the Gaussian bandit where the expectation in (2.2.2) can be computed explicitly. This explicit computation results in the similar expression to the UCB algorithm (2.2.1). In general, one may use an appropriate quadrature or Monte-Carlo approach to compute the expectation; we will use a Monte-Carlo approach for our simulations.

2.2.5 Thompson Sampling (TS)

Thompson sampling [104] was the first proposed algorithm to tackle the multi-armed bandit problem. The idea of the algorithm is to sample $\hat{\Theta}_{t-1}$ from its current posterior distribution. The (random) decision at time t can then be given by

$$A_t^{TS} = \underset{k=1,\dots,K}{\arg\max}\, \mathbb{E}[R^{(k)}(Y^{(k)})|\Theta = \hat{\Theta}_{t-1}].$$

An overview of the practical aspects of Thompson sampling, including a general approach to obtain posterior samples, can be found in Russo et al. [98].

Thompson sampling was largely ignored in the academic literature for the first eight decades after it was discovered. This is partly due to the lack of mathematical guarantees for the algorithm. In recent decades, Thompson sampling was found to perform empirically well and was brought to the attention of theoreticians. The first theoretical guarantee was proved by Agrawal and Goyal [2], regarding the frequentist regret for the Bernoulli bandit. Further references can be found in the bibliographic remarks of Lattimore and Szepesvári [72, Chapter 36].

Russo and Van Roy [96] use an information-theoretic analysis to prove Bayesian regret bounds for Thompson Sampling, for many variations of the bandit problem. They show that Thompson Sampling achieves an optimal or near-optimal regret rate in those cases.

2.2.6 Information-Directed Sampling (IDS)

In an information-theoretic analysis of Thompson Sampling, Russo and Van Roy [96] establish an upper bound for the Bayesian regret by considering an information ratio,

$$\Psi_{t-1}^U(u) := \frac{\Delta_{t-1}(u)^2}{G_{t-1}(u)} \qquad ; u \in \Delta^K,$$

where

$$\Delta_{t-1}^U(u) := \sum_{k\in[K]} u_k \mathbb{E}\Big(R^{(A^*)}(Y_t^{(A^*)}) - R^{(k)}\big(Y_t^{(k)}\big)\Big|\mathcal{F}_{t-1}^U\Big),$$

$$G_{t-1}^U(u) := \sum_{k\in[K]} u_k \mathbb{E}\Big(\mathcal{H}_t^U(A^*) - \mathcal{H}_{t-1}^U(A^*)\big)\Big|\mathcal{F}_{t-1}^U, A_t = k\Big),$$

$A^* := \arg\max_{k\in[K]} \mathbb{E}\big(R^{(k)}(Y^{(k)})\big|\Theta\big)$ and $\mathcal{H}_t^U(X)$ is the Shannon entropy of a random variable X conditional on $\mathcal{F}_t^U$.

Russo and Van Roy [96] show that for any strategy U, a bound on the Bayesian regret can be given by

$$\mathcal{BR}(U,T,\pi) \le \sqrt{T\bar{\Psi}_T^U \log K} \quad \text{where} \quad \bar{\Psi}_T^U := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\pi\big(\Psi_t^U(U_t)\big).$$

Thus, establishing the regret bound is simplified to establishing a uniform bound on $\mathbb{E}_\pi\big(\Psi_t^{U^{TS}}(U_t^{TS})\big)$.

This method of proving the regret bound inspired them to propose another algorithm, called Information-Directed Sampling (IDS) [97]. The probabilistic decision of IDS is defined recursively by

$$U_t^{IDS} := \arg\min_{u\in\Delta^K} \Psi_{t-1}^{U^{IDS}}(u) \tag{2.2.3}$$

This approach gives a direct method to deal with extra information that some algorithms cannot deal with (see Section 3.1 for further discussion).

Since $\mathbb{E}_\pi\big(\Psi_t^{U^{IDS}}(U_t^{IDS})\big) \le \mathbb{E}_\pi\big(\Psi_t^{U^{TS}}(U_t^{TS})\big)$, all established bounds for Thompson Sampling [96] also holds for the IDS. However, there are a couple of challenges in applying the IDS in practice. The first is to solve the K-dimensional optimisation (2.2.3). Russo and Van Roy [97] show that the solution to (2.2.3) can be achieved by a sparse U. In particular, we can choose U^{IDS} such that $\#\{k : U_t^{(k),IDS} > 0\} \le 2$. Hence, (2.2.3) can be solved by applying a simple search.

The second difficulty for the IDS is to compute G_t^U. One possible approach is to use Monte-Carlo simulation, but this is costly. Kirschner and Krause [67] consider a modification of the original IDS. In the Gaussian case, their proposed modification is equivalence to considering $\mathcal{H}_t^U(\Theta)$ instead of $\mathcal{H}_t^U(A^*)$. For simplicity in our implementation, we will consider the IDS algorithm using $\mathcal{H}_t^U(\Theta)$.

Chapter 3

Asymptotic Randomised Control

In Chapter 2, we discussed the optimal solution to the total discounted payoff for the classical bandit problem in terms of Gittins index. We also see in the approximation of the index with Gaussian reward that the index can be decomposed into two components as in the UCB algorithm:

$$\alpha = \text{Exploitation gain} + \text{Learning premium}, \tag{3.0.1}$$

where the 'Learning premium' increases with uncertainty of our estimate.

Gittins' theorem shows the optimality of index strategy for the classical bandits where the arms are independent. The optimal strategy for dependent arms have not yet been studied. We may also expect the (near) optimal strategy for dependent arms to be an index strategy of the form 3.0.1. The question is how shall we choose such index?

In the UCB algorithm (Section 2.2.3), one often considers index-based strategy of the form 3.0.1 where the 'Learning premium' is only quantified through the uncertainty of the estimated reward of the chosen arm. Is the uncertainty of the chosen arm is an appropriate choice of 'Learning premium' with correlated information?

To answer this question, we consider a small variation of the classical bandit. Suppose we have an arm that gives a reward from $N(\Theta, 1)$ where after plays, we have a posterior $\Theta \sim N(m, d)$. Now, suppose the second arm always gives a constant reward of £1, but it also provides an observation from $N(\Theta, 1)$ without affecting the reward.

When ignoring the effect of uncertainty aversion, it is more reasonable to play the second arm when $m < 1$. This means that if we use an index-based policy for our decision, the first arm's learning premium should be added to the second arm's reward, which is £1. The second arm itself does not have uncertainty in its reward.

In particular, we can argue from this observation that it is not the uncertainty of the reward that we wish to take as a learning reward, but it is the information that the arm can provide.

In this chapter, we will come up with an appropriate choice of (semi-) index by considering the dynamic programming as in (2.1.2) and add a diversity premium to smooth the value function. The learning premium of this (semi-) index strategy arrives naturally from the diffusive change in an approximate value function. This premium quantifies the value of information provided by each arm. Hence, we obtain a strategy which has a natural interpretation to make decisions under a learning environment. In addition to the natural interpretation, we find that the resulting strategy does not necessarily suffer from the various theoretical limitations common to other bandit algorithms. We will begin this chapter by discussing those limitations and then introducing our approach.

3.1 Limitation of existing bandit algorithms

We have already discussed various approaches to tackle the multi-armed bandit problem in Chapter 2. These approaches often come from a heuristic strategy and prove that it works by considering a theoretical regret guarantee in a specific setting. However, these analogies could result in an ill decision in some other instances. This section will illustrate some examples when the problem arrives; many more examples can be found in Russo and Van Roy [97].

3.1.1 Incomplete Learning of deterministic policy

One may notice that many bandit algorithms, e.g. Gittins index, greedy algorithm (with $\epsilon = 0$), UCB algorithm and Knowledge Gradient, are

1. **Deterministic:** We can explicitly write down the decision given the past observation, i.e. A_t is $\mathcal{F}^U_{t-1}$-measurable rather than $\sigma(\mathcal{F}^U_{t-1}, \zeta_t)$-measurable, equivalently, U_t takes value in a standard basis.

2. **Time-independent:** The decision is made based on only the historical observed data, i.e. if $\mathcal{F}^U_t = \mathcal{F}^U_s$, then $U_{t+1} = U_{s+1}$.

This decision structure could result in the situation where the best option has not been played due to incomplete learning.

Suppose that the bandit has 2 arms. The first arm always gives a fixed reward, whereas the second arm's reward is generated from an unknown distribution. For any strategies that satisfy both properties above, if this strategy decides to play the first arm, it will never play the second arm again. However, we can see that if the mean reward of the first arm has an unbounded support, the probability that the first arm is better is strictly positive. This probability never changes when the first arm is abandoned. This means that we have a strictly positive probability of always playing sub-optimal options. [1]

3.1.2 Information Ignorance

Many works on bandits assume that all arms have an identical structure. Therefore, they fail to capture the setting where each arm provides a different information structure.

Suppose that the bandit has 100 arms. The first arm always gives a reward of 0, but it will tell us the parameters of all other arms. Suppose that other arms always give a strictly positive reward from an unknown distribution where its parameter will be revealed when the first arm is played.

In this circumstance, the first arm never has the best reward. Therefore, we can show that the first arm will never be played by the greedy algorithm (with $\epsilon = 0$), Thompson, or UCB algorithm. However, if we need to repeatedly make decisions in this system for a long time, it is probably worth playing the first arm once to learn the system and then playing the best arm afterward. [2]

This phenomenon shows that many algorithms for bandit fail to process information correctly as they are based on the assumption that every arm has the same information structure. In fact, Russo and Van Roy [97, Example 2] also give an example when each arm has the same structure, but we still see the failure of many classical approaches to choose decisions wisely.

3.1.3 Time Ignorance

One can see that apart from the IDS, other approaches described in Chapter 2 face either incomplete learning or the information ignorance effect. This is why Russo and Van Roy [97] propose to consider the IDS strategy instead.

[1] This sub-optimal consequence also appears for Gittins index even if this is an optimal solution to the total discounted problem. The reason is that Gittins is optimal for the discounted payoff. However, it is not generally optimal for the payoff with no discount factor.

[2] Vice versa, we will play the first arm too many times under the ETC or ϵ-greedy algorithm when $\epsilon > 0$.

Nonetheless, the IDS also has some flaws in its decisions as it does not take into account how much time is available. The main benefit of the IDS is that it prefers to make a non-profitable decision early on, to gain the information for the future. However, if we only play for a short period, the IDS still gives the same decision, and thus resulting in excessive information without reward.

Outline of the chapter

We have discussed some flaws in the existing algorithms that have been considered for the bandit problem. The IDS addresses incomplete learning and information ignorance but still suffers from the time ignorance effect. In contrast, the knowledge gradient method only faces incomplete learning, but it still captures the valuation of information via the reward's improvement (through the term insides expectation of (2.2.2)). Time availability can also be calibrated through the discount factor β.

Due to this observation, a heuristic solution to overcome these flaws is to consider a knowledge gradient method and enforce randomness in its decision. This chapter will more or less consider this approach but in more analytical detail. The extra randomness is added by introducing some diversity terms to our reward, allowing us to formulate (1.4.1) in terms of the relaxed control problem. We then solve this problem by approximating the value function directly rather than using the heuristic one-step look ahead argument as in the knowledge gradient method.

Formally, we will consider bandit problems in a Bayesian framework. We observe that when the prior distribution and the new observation form a prior-posterior conjugate pair, we can describe bandit problems as a classical stochastic control problem with a Markovian underlying state with a known dynamic. The underlying state only evolves with a small amount when our estimate is precise. We exploit this property to approximate the regularised finite horizon value function via a careful analysis of Taylor's expansion. We then take the limit on the horizon and obtain a semi-closed form approximation to the value function and its near-optimal control. Finally, we study the trade-off between the entropy regularisation coefficient and errors of our approximation. We then establish a scheme to construct an algorithm for bandit problems.

The chapter proceeds as follows. In section 3.2, we state a generic control problem with an underlying Markov process and state relevant assumptions. We then considered a few examples of the bandit problems in 3.3, and showed that those examples can be formulated in our considered framework and satisfy the required assumptions. Section 3.4 and 3.5 derive a formal approximation to our control problem. Most

proofs of interims results are simple but rather tedious. We thus provide them in Appendix A.

This chapter is based on the paper [34].

3.2 An Encompassing Stochastic Control Formulation

In this section, we will state our problem setup and assumptions that we consider throughout this chapter. The stated problem is a classical control problem in discrete time with an underlying Markov process. The stated assumption is inspired by bandit problems where we will discuss the connection between our problem setup and bandit problems in Section 3.3.

Let $(\Omega, \mathbb{P}, \mathcal{F})$ be a probability space equipped with $(Z_t)_{t\in\mathbb{N}}$, a sequence of IID F-dimensional random variables with $\mathbb{E}(Z_t) = 0$, $\mathrm{Var}(Z_t) = I_F$, $\mathbb{E}|Z_t|^3 < \infty$, and a sequence $(\zeta_t)_{t\in\mathbb{N}}$ with $\zeta_t \sim_{IID} U[0,1]$ independent of (Z_t).

We define the filtration $\mathcal{F}_t := \sigma(Z_s, \zeta_s : s \le t)$. Again, (ζ_t) represents a random seed used to select a random decision in each time step, whereas (Z_t) represents the randomness of the outcome.

We also define the *random action* (A_t) and the *probabilistic decision* (U_t) as in Section 1.1.1 with $\mathcal{F}_t^U$ replaced by $\mathcal{F}_t$.

We model the state of our system (e.g. the posterior distribution of parameters to be estimated, the restless state, or the context of the bandit payoffs) by a Markov process X. (We will later decompose X to two components (M, D) to represent estimators and their precision.) Suppose that, when the control is chosen, we obtain a reward and X evolves corresponding to the choice of control. In particular, when we choose $i \in [K] := \{1, ..., K\}$, the underlying state evolves according to the transition map

$$\Phi(x, i, z) = x + b_\Phi^{(i)}(x) + \sigma_\Phi^{(i)}(x)z \qquad : \qquad x \in \mathbb{R}^L, i \in [K], z \in \mathbb{R}^F \tag{3.2.1}$$

where for each $i \in [K]$, $b_\Phi^{(i)} : \mathbb{R}^L \to \mathbb{R}^L$ and $\sigma_\Phi^{(i)} : \mathbb{R}^L \to \mathbb{R}^{L\times F}$.

Given a control $U \equiv (U_1, U_2, ...) \in \mathcal{U}$, we denote by $X^{x,U}$ the following discrete time dynamics: $X_0^{x,U} = x$ and $X_{t+1}^{x,U} = \Phi(X_t^{x,U}, A(U_{t+1}, \zeta_{t+1}), Z_{t+1})$ for $t \ge 0$.

Let the expected payoff of an action $i \in [K]$, conditional on the state $x \in \mathbb{R}^L$, be denoted by $f : \mathbb{R}^L \times [K] \to \mathbb{R}$ and the discount rate be $\beta \in (0, 1)$. The objective of our problem is to solve the optimisation

$$V(x) := \sup_{U\in\mathcal{U}} \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t \Big(f\big(X_t^{x,U}, A(U_{t+1}, \zeta_{t+1})\big)\Big)\Big]. \tag{3.2.2}$$

In the context of bandit problems, $f(x, i)$ shall be interpreted as the expected reward when we choose the ith option and parameters of the posterior distribution (and contextual elements) are given in form of x.

Remark 3.1. As discussed in Section 1.6, we will interpret (X_t^U) as a posterior parameter. In general, one may simply remove the IID assumption on (Z_t) and allow the distribution of Z_t to vary with the control and the observation up to time t. In this case, we require that (Z_t) is a martingale difference process with a (conditional) unit variance and uniformly bounded (conditional) third absolute moment. However, a careful construction of filtrations (see Section 3.3) and a consideration of the dynamic programming principle (Theorem 3.4) are required. The rest of the computation remains the same. We will restrict our attention to the IID case, to simplify our discussion.

We aim to give an approximation to (3.2.2) for a bandit problem by introducing a diversity premium. We will state the required assumptions for ease of reference here, and we will justify them in Section 3.3.

3.2.1 Notation

We first introduce the notation that will be used in our discussion.

- The superscript (i) should be interpreted as indicating an association with an action $i \in [K]$. A subscript will identify the components of vectors, matrices, or tensors of interest. Throughout this chapter, it may also be the case that the ith component of a vector is associated with the ith action. In that case, we may use either sub- or superscripts to simplify notation.

- For any function $h : \mathbb{R}^G \times \mathbb{R}^H \to \mathbb{R}^C$, we write $\partial_m^k \partial_d^l h$ for a tensor of degree $(k, l, 1)$ (i.e. taking values in $G^k H^l C$-dimensional space) corresponding to the kth derivative with respect to m and lth derivative with respect to d of the function h.

- Let P and Q be I-dimensional tensors of degree k. We write

$$\langle P; Q \rangle := \sum_{i_1,\ldots,i_k=1}^{I} P_{i_1,\ldots,i_k} Q_{i_1,\ldots,i_k}.$$

  In particular, if P and Q are vectors, $\langle P; Q \rangle$ is a standard inner product, and if P and Q are matrices, then $\langle P; Q \rangle = \mathrm{Tr}(PQ)$. We write $|\cdot|$ for the corresponding (Euclidean) norm.

- We will use the families of polynomials

$$\mathcal{P}_r^+ := \{bx^r(1+x^n) : b>0, n\in\mathbb{N}\} \qquad \text{and} \qquad \mathcal{P}_r^- := \{b(1+x^{-r}) : b>0\}.$$

  To avoid confusion later, we specify that the coefficient b and the exponent n depend only on the constant C in Assumption 3.2 and Assumption 3.3 and will not depend on λ, a, (m,d), h and β (to be defined in the following sections).

Remark 3.2. We will use elements in $\mathcal{P}_r^+$ and $\mathcal{P}_r^-$ to track the error of our approximation. One more commonly uses big-$\mathcal{O}$ notation, however, in our analysis, we need to carefully consider the trade-off between the magnitude of different terms. For this reason, the use of $\mathcal{P}_r^+$ and $\mathcal{P}_r^-$ leads to better notational simplicity.

3.2.2 Assumptions

We now state the assumptions that will be used for our approximation in the multi-armed bandit problem.

Assumption 3.1. *The transition map* (3.2.1) *for the underlying process* $X=(M,D)$ *can be written in the form*

$$\begin{aligned}\Phi:\ \mathbb{R}^G\times\mathbb{R}^H\times[K]\times\mathbb{R}^F &\to \mathbb{R}^G\times\mathbb{R}^H\\ (m,d,i,z) &\mapsto \begin{pmatrix}m\\d\end{pmatrix}+\begin{pmatrix}\mu^{(i)}(m,d)\\ b^{(i)}(m,d)\end{pmatrix}+\begin{pmatrix}\sigma^{(i)}(m,d)z\\0\end{pmatrix}.\end{aligned}$$

where $\mu^{(i)}, b^{(i)}$ *and* $\sigma^{(i)}$ *are functions from* $\mathbb{R}^G\times\mathbb{R}^H$ *to a space with an appropriate dimension.*

Assumption 3.2. *There exists a constant* $C<\infty$ *such that for all functions* $\psi\in\left\{b^{(i)},\mu^{(i)},\left(\sigma^{(i)}(\sigma^{(i)})^\top\right): i\in[K]\right\}$*, we have*

$$\sup_m|\psi(m,d)|\le C|d|^2,\quad \sup_m|\,\partial_m\psi(m,d)|\le C|d|^2,\qquad \text{and}\qquad \sup_m|\,\partial_d\psi(m,d)|\le C|d|.$$

Remark 3.3. These assumptions are general enough to include, for example, the (extended) Kalman filter dynamics, the restless state or some small contextual elements for the underlying state.

We usually view m as an estimator of the parameters in our model, while d represents the variance, or *inverse precision*, of our estimates. (The decomposition of the underlying state considered here is referred to as the 'knowledge state' in the Knowledge Gradient Algorithm considered by Ryzhov et al. [100].)

For simplicity in our analysis, we will also assume that the instantaneous reward function f is uniformly bounded and its derivatives are also uniformly bounded.

Assumption 3.3. *Write* $f : \mathbb{R}^G \times \mathbb{R}^H \to \mathbb{R}^K$ *for the vector function with components* $f_i(m,d) = f^{(i)}(m,d) = f(m,d,i)$. *We assume that* $f \in C_b^3(\mathbb{R}^G \times \mathbb{R}^H; \mathbb{R}^K)$.

3.3 Multi-armed bandit as a Control problem

The key difference between the multi-armed bandit and the stochastic control problem discussed in Section 3.2 is the interaction between controls and filtrations. For the bandit problem, our control determines the observation (and thus the filtration) whereas this is not the case for the classical control problem where the filtration is predetermined.

This section will discuss how to formulate some examples of multi-armed bandit problems in terms of a classical control proposed in the earlier section by seeing bandits through the Bayesian framework. We will see bandit problems as a Kalman filter and obtain our fixed observation as our innovation process. Furthermore, we will see the corresponding dynamic of these examples satisfying the assumptions made in Section 3.2.2.

3.3.1 Multi-armed bandit with additional information

We will start our discussion by generalising a classical multi-armed bandit. This model assumes that the reward of each bandit associate to a multi-dimensional parameter. When we play an arm, such arm produces a reward sampled from a distribution parameterised by such parameter. We assume that each component of the parameter under the prior is independent and we assume that we obtain observations regarding each component explicitly. We will later extend this to a more general setting in Section 3.3.2

Suppose that we choose from K arms of a bandit. When the ith arm is chosen, we observe the random variables $Y^{(i,j)} \sim N(\Theta^{(j)}, 1/p_{ij})$ for $i,j = 1,2,...,K$ chosen independently between each trial. The parameters $(\Theta^{(i)})$ are unknown parameters whereas the (p_{ij}) are known. We write $Y^{(i)} := \left(Y^{(i,j)}\right)_{j=1,...,K}$. The reward when the ith arm is chosen is given by $R^{(i)}(Y^{(i)})$ where $R^{(i)} : \mathbb{R}^K \to \mathbb{R}$ is a known bounded function. (N.B. In the classical bandit problem, rewards of each arm depends only on each component of the parameter. Moreover, the reward is the only observation. In this case, we have $p_{ij} = 0$ if $i \neq j$ and we have $R^{(i)}(y)$ as a function which only depends on the ith component of y.)

Remark 3.4. By considering the posterior update, one may interpret p_{ij} as a proxy for 'the number of observations' obtained from the jth arm when the ith arm is chosen

(even if p_{ij} might be non-integer). We can also allow, formally, $p_{ij} = 0$ for some $j \neq i$ to indicate that playing the ith arm does not give information about the jth arm, in this case, we set $Y^{(i,j)} = 0$ for simplicity.

We recall that the objective of the multi-armed bandit problem is to solve

$$V(m,d) = \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\left[\sum_{t=0}^{\infty} \beta^t R^{(A_{t+1})}\left(Y_{t+1}^{(A_{t+1})}\right)\right]. \tag{3.3.1}$$

where $A_t = A(U_t, \zeta_t) = \inf\left\{i : \sum_{k=1}^{i} U_t^{(k)} \geq \zeta_t\right\}$ and $N(m, \mathrm{Diag}(d))$ is the prior of Θ at time 0.

Recall that the filtration of the multi-armed bandit problem is then given by $\mathcal{F}_t^U := \sigma(\zeta_1, Y_1^{(A_1)}, ..., \zeta_t, Y_t^{(A_1)})$ where U is our chosen probabilistic decision and ζ_t is a random element as described in Section (1.1.1).

Now, assume that the prior, at time t, of the parameter Θ, is given by $\mathrm{Law}(\Theta|\mathcal{F}_t^U) = N(M_t, \mathrm{Diag}(D_t))$. Then the posterior distribution after observing $Y_{t+1}^{(i,j)}$ can be given by $\mathrm{Law}(\Theta^{(j)}\Big|\mathcal{F}_t^U, Y_{t+1}^{(i,j)}) = N\Big(M_{t+1}^{(j)}, D_{t+1}^{(j)}\Big)$, where

$$M_{t+1}^{(j)} = \frac{\left(D_t^{(j)}\right)^{-1} M_t^{(j)} + p_{ij} Y_{t+1}^{(i,j)}}{\left(D_t^{(j)}\right)^{-1} + p_{ij}} \qquad \text{and} \qquad \frac{1}{D_{t+1}^{(j)}} = \frac{1}{D_t^{(j)}} + p_{ij}. \tag{3.3.2}$$

One can show that $\mathrm{Law}(Y_{t+1}^{(i,j)}\Big|\mathcal{F}_t^U) = N\Big(M_t^{(j)}, D_t^{(j)} + 1/p_{ij}\Big)$. Hence, we can write

$$\left.\begin{aligned} M_{t+1}^{(j)} &= M_t^{(j)} + D_t^{(j)}\left(\frac{p_{ij}}{1 + D_t^{(j)} p_{ij}}\right)^{1/2} Z_{t+1}^{(i,j)}, \\ D_{t+1}^{(j)} &= D_t^{(j)} - \left(D_t^{(j)}\right)^2\left(\frac{p_{ij}}{1 + D_t^{(j)} p_{ij}}\right), \end{aligned}\right\} \tag{3.3.3}$$

where $Z_{t+1}^{(i,j)} = \left(D_t^{(j)} + \frac{1}{p_{ij}}\right)^{-1/2}\left(Y_{t+1}^{(i,j)} - M_t^{(j)}\right)$. We then have $Z_{t+1}^{(i,j)} \sim N(0,1)$ for $p_{ij} \neq 0$. Moreover, one can show by the tower property that, if $f(m,d,i) := \mathbb{E}\left[R^{(i)}(Y^{(i)})\Big|\Theta \sim N(m, \mathrm{Diag}(d))\right]$ denotes the one-step conditional expected payoff,

$$\mathbb{E}_{m,d}\left[R^{(A_{t+1})}\left(Y_{t+1}^{(A_{t+1})}\right)\right] = \mathbb{E}_{m,d}\left[f\left(M_t^{m,d,U}, D_t^{m,d,U}, A_{t+1}\right)\right].$$

Hence, from the bounded convergence theorem, (3.3.1) becomes

$$V(m,d) = \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\left[\sum_{t=0}^{\infty} \beta^t f\left(M_t^{m,d,U}, D_t^{m,d,U}, A_{t+1}\right)\right]. \tag{3.3.4}$$

As the distribution $Z^{(i,j)}$ does not depend on i, one can see (3.3.4) and (3.3.3) as a classical control problem, as discussed in Section 3.2, with drifts and volatilities

$$\mu^{(i)}(m,d) = 0, \qquad b^{(i)}(m,d) = \left(-d_1^2\left(\frac{p_{i1}}{1+d_1 p_{i1}}\right), ..., -d_K^2\left(\frac{p_{iK}}{1+d_K p_{iK}}\right)\right) \qquad \text{and}$$

$$\sigma^{(i)}(m,d) = \mathrm{Diag}\left(d_1\left(\frac{p_{i1}}{1+d_1 p_{i1}}\right)^{1/2}, ..., d_K\left(\frac{p_{iK}}{1+d_K p_{iK}}\right)^{1/2}\right),$$

which satisfies Assumption 3.1 and 3.2.

Remark 3.5. In our problem formulation, we assume that the filtration is generated by the processes (Z_t) (and (ζ_t)). In the context of the multi-armed bandit problem, we see that our action determine the filtration. However, as discussed in the earlier paragraph, the distribution of $Z^{(i,j)}$ does not depend on i. Hence, we can treat $Z^{(i,j)}$ as our fixed observation which evolves the underlying state (which is the posterior parameter). It is worth pointing out that, we do not observe the exact process (Z_t) but instead, conditional on $A_t = i$, we obtain the observation $\left(Y_{t+1}^{(i,j)}\right)_{j:\, p_{ij}\neq 0}$ when the ith arm is chosen. We thus define (Z_t) to be the *innovation process* corresponding to our observation, that is,

$$Z_{t+1}^{(j)} \begin{cases} = \left(D_t^{(j)} + 1/p_{ij}\right)^{1/2}\left(Y_{t+1}^{(i,j)} - M_t^{(j)}\right) & : p_{ij} \neq 0, \\ \sim N(0,1) & : p_{ij} = 0. \end{cases}$$

The additional randomness for j such that $p_{ij} = 0$ is considered to ensure that the bandit problem is well-aligned with the original setting. However, these additional random variables $(Z_t^{(j)})_{j:p_{ij}\neq 0}$ do not contribute to the problem, as $\sum_{j:p_{ij}=0} \sigma_{kj}^{(i)}(m,d) z_j = 0$ for all $k = 1, ..., K$.

3.3.2 Correlated bandit

In Section 3.3.1, we considered the model where we have separate observations to infer each component of the parameter Θ. However, this may not be the case in general; for example, our observations $Y^{(i)}$ may be sampled from a distribution with a parameter $\Gamma^{(i)}$ where $\Gamma^{(i)}$ is a (linear) combination of the parameter Θ. In particular, we receive an information in terms of 'correlation'.

Suppose that we have an unknown parameter Θ taking values in $\mathbb{R}^G$. Suppose that when the ith option is chosen, we obtain an observation $Y^{(i)} \sim N(b_i^\top \Theta, P_i^{-1})$, where $b_i \in \mathbb{R}^{G\times l}$ and P_i is a positive definite matrix of dimension $l \times l$ and obtain an instantaneous reward $R^{(i)}(Y^{(i)})$.

Now suppose that the prior at time t of the parameter Θ is given by $\text{Law}(\Theta|\mathcal{F}_t^U) = N\big(M_t, \Sigma_t\big)$. The conditional density of Θ given $Y_{t+1}^{(i)}$ and $\mathcal{F}_t^U$ is

$$\phi(\theta|\mathcal{F}_t^U, Y_{t+1}^{(i)}) \propto \exp\bigg(-\frac{1}{2}\theta^\top\big(\Sigma_t^{-1} + b_i P_i b_i^\top\big)\theta + \theta^\top\big(\Sigma_t^{-1}M_t + b_i P_i Y_{t+1}^{(i)}\big)\bigg).$$

In particular, after observing $Y_{t+1}^{(i)}$, the posterior distribution is $\text{Law}(\Theta|\mathcal{F}_t^U, Y_{t+1}^{(i)}) = N\big(M_{t+1}, \Sigma_{t+1}\big)$[3], where

$$M_{t+1} = \big(\Sigma_t^{-1} + b_i P_i b_i^\top\big)^{-1}\Big(\Sigma_t^{-1}M_t + b_i P_i Y_{t+1}^{(i)}\Big), \quad \text{and} \quad \Sigma_{t+1} = \big(\Sigma_t^{-1} + b_i P_i b_i^\top\big)^{-1}. \tag{3.3.5}$$

Due to the recursive definition of Σ_t, we write $\Sigma_t^{-1} = \Sigma_0^{-1} + \sum_{k=1}^K N_t^{(k)}\big(b_k P_k b_k^\top\big)$ where $N_t^{(i)}$ is the number of times that the ith arm was chosen.

Writing $D_t^{(i)} = 1/N_t^{(i)}$, we represent the posterior variance Σ_t by $\Sigma_t = \Sigma(D_t)$ where

$$\Sigma(d) := \bigg(\Sigma_0^{-1} + \frac{1}{d_1}\big(b_1 P_1 b_1^\top\big) + \ldots + \frac{1}{d_K}\big(b_K P_K b_K^\top\big)\bigg)^{-1}.$$

and Σ_0 is chosen to be the prior variance of Θ.

One can also show that the distribution of $Y_{t+1}^{(i)}|\mathcal{F}_t^U$ is $N\Big(b_i^\top M_t, \Big(b_i^\top\Sigma(D_t)b_i + P_i^{-1}\Big)\Big)$. Hence, we can represent (3.3.5) by

$$\left.\begin{aligned} M_{t+1} &= M_t + \big(\Sigma(D_t)^{-1} + b_i P_i b_i^\top\big)^{-1} b_i P_i\Big(b_i^\top\Sigma(D_t)b_i + P_i^{-1}\Big)^{1/2} Z_{t+1}^{(i)}, \\ D_{t+1}^{(j)} &= D_t^{(j)} + \left(\frac{-\big(D_t^{(i)}\big)^2}{1 + D_t^{(i)}}\right)\mathbb{I}(i=j) \end{aligned}\right\} \tag{3.3.6}$$

where the square root of the matrix may be chosen such that its eigenvalues are non-negative.

As in Subsection 3.3.1, the objective of the linear bandit problem becomes

$$V(m,d) = \sup_{U\in\mathcal{U}}\mathbb{E}\bigg[\sum_{t=0}^{\infty}\beta^t f\big(M_t^{m,d,U}, D_t^{m,d,U}, A_{t+1}\big)\bigg], \tag{3.3.7}$$

where f is the the one-step conditional expected payoff given by

$$f(m,d,i) = \int_{\mathbb{R}^l} R^{(i)}\Big(b_i^\top m + \Big(b_i^\top\Sigma(d)b_i + P_i^{-1}\Big)^{1/2} z\Big)\varphi_l(z)dz$$

and φ_l is the density function of $N(0, I_l)$.

[3]One may see the posterior update as a Kalman filtering problem where the parameter Θ is a (static) hidden state.

Remark 3.6. Rusmevichientong, Mersereau and Tsitsiklis [93] consider the one dimensional version ($l = 1$) of this problem when $R^{(i)}(y) = y + \psi_i$ and $b_i \neq 0$ for all i. They show that, in this case, the 'Greedy algorithm' (i.e. an algorithm which always chooses an option maximising $f(m, d, i) = b_i m + \psi_i$) eventually coincides with an optimal solution to (3.3.7). Moreover, this algorithm achieves a regret of order $\mathcal{O}(\log T)$. In the later section, we will see that our learning premium converges to zero as $|d| \to 0$ and $|D_t| \to 0$ a.s. Hence, one may extend analysis of [93] to establish a regret bound for this simple case. Nonetheless, in this thesis, we will focus on the solution to the control problem. We will leave the regret analysis as our future work.

Remark 3.7. The choice of d described above ensure that $t \mapsto |D_t|$ is decreasing. In fact, one can consider defining d to be the covariance matrix Σ directly. All of the followed-up analysis holds but we must considered an appropriate norm (e.g. an operator norm) to ensure that $t \mapsto |\Sigma_t|$ is decreasing. Then the stated result follows by using the equivalence of norm in the finite dimensional space.

3.4 Approximation of the value function

In Section 3.3, we discussed how a bandit problem could be seen as a classical stochastic control problem. We see in (3.3.3) and (3.3.6) that the corresponding underlying state in the stochastic control problem can be decomposed into two components, (M_t) and (D_t). We can verify that the evolution of these components satisfies Assumptions 3.1, 3.2 and 3.3.

Throughout the rest of this chapter, we will focus on the classical control problem stated in Section 3.2 with the stated assumptions (inspired by bandit problems).

Recall that our objective is to give an approximation to the problem

$$V(m, d) = \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\left[\sum_{t=0}^{\infty} \beta^t f\left(M_t^U, D_t^U, A_{t+1}\right)\right] \qquad \text{where} \quad A_t = A(U_t, \zeta_t). \tag{3.4.1}$$

By writing down the Bellman equation, we can see that the optimal randomised policy for the above problem is often a Dirac measure, i.e., an optimal decision will always choose a single action with a probability 1. Consequently, the optimal control involves the argmax function, which is, in general, non-smooth and can lead to sensitivity of the control to perturbation of the coefficients (b, μ, σ, f) as discussed in Reisinger and Zhang [88].

To avoid Dirac-like controls, in this section, we modify the value function (3.4.1) by introducing a regularisation to smooth the value function. We then proceed by

approximating the smooth finite horizon value function and then allowing the horizon to tend to infinity. We will continue our analysis further in Section 3.5 to remove the regularisation and show that our approximation method gives a sensible solution to the stated control problem.

3.4.1 Relaxed control problem

We add a diversity premium to our payoff to encourage randomised strategies and give smoothness to our value function. This reward ensures that the optimal control U takes values in the relative interior of Δ^K; in particular, it means that a random decision is always preferred.

Definition 3.1. Let $\mathcal{H} : \Delta^K \to \mathbb{R}$ be a smooth entropy function (see Definition 3.3) and $\lambda > 0$. We define the regularised value function to be

$$V_\infty^\lambda(m,d) \coloneqq \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t \Big(f\big(M_t^U, D_t^U, A_{t+1}\big) + \lambda \mathcal{H}(U_{t+1}) \Big) \Big] \tag{3.4.2}$$

where $A_t = A(U_t, \zeta_t)$.

In this section, we will derive an approximation of the form

$$V_\infty^\lambda(m,d) \approx \Big(\frac{1}{1-\beta}\Big)\big(\mathrm{S}_{\max}^\lambda \circ \alpha^\lambda\big)(m,d). \tag{3.4.3}$$

The function $\mathrm{S}_{\max}^\lambda : \mathbb{R}^K \to \mathbb{R}$ is a soft version of the maximum function, related to the choice of smooth entropy $\mathcal{H}$. The function α^λ is a limit of recursive functions, which approximates the incremental value of a single step which can be seen as a (semi-) index for our decision. This section is devoted to deriving (3.4.3) and quantifying its error.

We will first introduce the ingredients for our approximation.

3.4.2 Smooth entropy and related functions

Smooth Entropy

In order to obtain an approximation to the value function in (3.4.3), we want to choose a smooth entropy function $\mathcal{H}$ such that

$$\mathrm{S}_{\max}^\lambda(a) \coloneqq \sup_{u \in \Delta^K} \Big(\sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big) \tag{3.4.4}$$

is 3-times differentiable and its derivative is a polynomial in $1/\lambda$. We may see $\mathrm{S}_{\max}^\lambda$ as a smooth approximation to the maximum function. This can be seen through the fact that $S_{\max}^0(a) = \max_i a_i$.

In order to do so, we observe that $\mathrm{S}^{\lambda}_{\max}(a) = \lambda S(a/\lambda)$, where

$$S(a) = \sup_{u \in \Delta^K} \Big(\sum_{i=1}^{K} u_i a_i + \mathcal{H}(u) \Big). \tag{3.4.5}$$

Hence, the function S can be interpreted as a convex conjugate of $-\mathcal{H}$ (see e.g. Rockafellar [92]).

The requirement of the differentiability of S restricts our choices of $\mathcal{H}$. Therefore, it may be easier to choose S first and define $\mathcal{H}$ by convex duality as a conjugate of S.

By considering the form of (3.4.5), we can see that S must satisfy the properties of a 'nonlinear expectation' (or equivalently 'risk measure') defined on a finite space (See Coquet et al. [37]). The term $\mathcal{H}$ can thus be seen as a penalty term in the robust representation of a nonlinear expectation.

Definition 3.2. We say a function $S : \mathbb{R}^K \to \mathbb{R}$ is a *convex nonlinear expectation* if it satisfies the following:

(i) **Monotonicity**: If $a \le b$, then $S(a) \le S(b)$;

(ii) **Translation Equivariance**: For all $c \in \mathbb{R}$, $S(a + c\mathbb{1}) = S(a) + c$;

(iii) **Convexity**: For any $\kappa \in [0, 1]$, $S(\kappa a + (1-\kappa)b) \le \kappa S(a) + (1-\kappa)S(b)$;

where the inequality above shall be interpreted component-wise and $\mathbb{1}$ is the vector with all entries 1.

One can see that if the function S admits a representation (3.4.5) then S must be a convex nonlinear expectation. By considering $\mathbb{R}^K$ as a space of random variables defined on a finite sample space, the converse follows from Föllmer and Schied [49, Theorem 4.16] (see also Frittelli and Rosazza Gianin [50]) with appropriate sign changes.

Theorem 3.1. *A convex nonlinear expectation S admits a representation of the form*

$$S(a) = \sup_{u \in \Delta^K} \Big(\sum_{i=1}^{K} u_i a_i + \mathcal{H}_{\max}(u) \Big),$$

where

$$\mathcal{H}_{\max}(u) := - \sup_{a \in \mathcal{A}_S} \Big(\sum_{i=1}^{K} u_i a_i \Big), \qquad \mathcal{A}_S := \{a \in \mathbb{R}^K : S(a) \le 0\}.$$

Furthermore, $\mathcal{H}_{\max}$ is the maximal function which represents S, i.e. if there exists $\mathcal{H}$ such that (3.4.5) *holds with $\mathcal{H}$, then $\mathcal{H}(u) \le \mathcal{H}_{\max}(u)$ for all $u \in \Delta^K$.*

In order to approximate the value function (3.4.3), we would like $\mathrm{S}^{\lambda}_{\max}(a)$ to be a (uniform) approximation to $\max_i a_i$. At the same time, we also would like to guarantee that $\mathcal{H}(u)$ is uniformly bounded, as this will help to ensure that the induced error in our value function (3.4.2) does not explode. In fact, these two properties are equivalent.

Theorem 3.2. *Let S be a convex nonlinear expectation. The following are equivalent.*

(i) There exists $N \in \mathbb{R}$ such that $S(a) + N \geq \max_i a_i$ for all $a \in \mathbb{R}^K$.

(ii) There exists $N \in \mathbb{R}$ such that $\mathcal{A}_S := \{a \in \mathbb{R}^K : S(a) \leq 0\} \subseteq (-\infty, N]^K$.

(iii) There exists a bounded function $\mathcal{H} : \Delta^K \to \mathbb{R}$ such that (3.4.5) holds.

(iv) For $\mathrm{S}^{\lambda}_{\max}(a) = \lambda S(a/\lambda)$, we have $\sup_{a \in \mathbb{R}^K} |\mathrm{S}^{\lambda}_{\max}(a) - \max_i a_i| \to 0$ as $\lambda \downarrow 0$.

Definition 3.3. We say a function $S : \mathbb{R}^K \to \mathbb{R}$ is a *smooth max approximator* if it is a 3-times differentiable convex nonlinear expectation with uniformly bounded derivatives such that Theorem 3.2 holds. We say a bounded function $\mathcal{H} : \Delta^K \to \mathbb{R}$ is a *smooth entropy* if it corresponds to the robust representation of a smooth max approximator.

Expansion terms

In the previous section, we have defined a smooth entropy and its convex conjugate, a smooth max approximator S. In our analysis, we will need to consider the derivatives of these functions.

Definition 3.4. Define

$$\nu^{\lambda}(a) := \partial_y S\Big|_{y=a/\lambda} \qquad \text{and} \qquad \eta^{\lambda}(a) = \partial_y^2 S\Big|_{y=a/\lambda}. \tag{3.4.6}$$

It follows that $\partial_a \mathrm{S}^{\lambda}_{\max}(a) = \nu^{\lambda}(a)$ and $\partial_a^2 \mathrm{S}^{\lambda}_{\max}(a) = \frac{1}{\lambda}\eta^{\lambda}(a)$.

Remark 3.8. If S admits the representation (3.4.5), then by Fenchel's inequality, it follows that

$$\nu^{\lambda}(a) = \arg\max_{u \in \Delta^K} \Big(\sum_{i=1}^{K} u_i a_i + \lambda \mathcal{H}(u) \Big).$$

In particular, we will see that ν^{λ} will be used as a proxy to compute our near-optimal control.

Example 3.1. A classical example for a smooth max approximator and its smooth entropy pair is a log-sum-exp function $S(a) \coloneqq \log\Big(\sum_{i=1}^{K} \exp(a_i)\Big)$. The corresponding entropy is given by the Shannon entropy $\mathcal{H} \coloneqq -\sum_{i=1}^{K} u_i \ln u_i$ with the convention that $0 \ln 0 = 0$. In this case, we can compute $\nu_i^\lambda(a) = \exp\big(a_i/\lambda\big)\Big/\Big(\sum_{j=1}^{K} \exp\big(a_j/\lambda\big)\Big)$ and $\eta_{ij}^\lambda(a) = \nu_i^\lambda(a)\big(\mathbb{I}(i=j) - \nu_j^\lambda(a)\big)$.

We also need to consider the derivatives of some composite functions.

Lemma 3.3. *Let $a(m,d) = f(m,d) + \phi$ where ϕ is constant in (m,d) be a function taking value in $\mathbb{R}^K$. Then*

$$\begin{aligned}
\partial_d\big(\mathrm{S}_{\max}^\lambda \circ a\big)(m,d) &= \mathcal{B}^\lambda\big(a(m,d),m,d\big),\\
\partial_m\big(\mathrm{S}_{\max}^\lambda \circ a\big)(m,d) &= \mathcal{M}^\lambda\big(a(m,d),m,d\big),\\
\partial_m^2\big(\mathrm{S}_{\max}^\lambda \circ a\big)(m,d) &= \Sigma^\lambda\big(a(m,d),m,d\big).
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{B}^\lambda\big(a,m,d\big) &= \sum_{j=1}^{K} \nu_j^\lambda(a)\, \partial_d f^{(j)}(m,d),\\
\mathcal{M}^\lambda\big(a,m,d\big) &= \sum_{j=1}^{K} \nu_j^\lambda(a)\, \partial_m f^{(j)}(m,d),\\
\Sigma^\lambda\big(a,m,d\big) &= \sum_{j=1}^{K} \Big(\nu_j^\lambda(a)\, \partial_m^2 f^{(j)}(m,d)\Big) + \frac{1}{\lambda}\sum_{i,j=1}^{K} \Big(\eta_{ij}^\lambda(a)\big(\partial_m f^{(i)}(m,d)\big)\big(\partial_m f^{(j)}(m,d)\big)^\top\Big)
\end{aligned}$$

In addition to these functions, we also introduce the following notation, which will prove useful in our approximation. We define $L^\lambda : \mathbb{R}^K \times \mathbb{R}^H \times \mathbb{R}^G \to \mathbb{R}^K$ by its components

$$\begin{aligned}
L_i^\lambda(a,m,d) \coloneqq \langle \mathcal{B}^\lambda(a,m,d); b^{(i)}(m,d)\rangle + \langle \mathcal{M}^\lambda(a,m,d); \mu^{(i)}(m,d)\rangle\\
+ \frac{1}{2}\langle \Sigma^\lambda(a,m,d); \sigma^{(i)}(m,d)\sigma^{(i)}(m,d)^\top\rangle.
\end{aligned} \tag{3.4.7}$$

Here, we may see L^λ as a proxy to the discrete time Hamiltonian.

We also recursively define $\alpha_T^\lambda, l_T^\lambda, F_T^\lambda : \mathbb{R}^H \times \mathbb{R}^G \to \mathbb{R}^K$ by $\alpha_1(m,d) = f(m,d)$ and

$$\left.\begin{aligned}
l_T^\lambda(m,d) &= L^\lambda\big(\alpha_T^\lambda(m,d),m,d\big),\\
F_T^\lambda(m,d) &= \beta l_{T-1}^\lambda(m,d) + ... + \beta^{T-1} l_1^\lambda(m,d),\\
\alpha_T^\lambda(m,d) &= f(m,d) + F_T^\lambda(m,d).
\end{aligned}\right\} \tag{3.4.8}$$

We will later see that those introduced functions will give an easy arithmetical method to approximate the value function when $|d|$ is small. Furthermore, these iteration leads to the solution to a fixed point problem.

3.4.3 Approximation to the horizon-T value function

In order to establish our approximation (3.4.3), we will consider the finite horizon problem,

$$V_T^{\lambda}(m,d) := \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\left[\sum_{t=0}^{T-1} \beta^t \Big(f\big(M_t^U, D_t^U, A_{t+1}\big) + \lambda \mathcal{H}(U_{t+1}) \Big) \right]. \tag{3.4.9}$$

We have a form of dynamic programming, written in terms of the horizon.

Theorem 3.4. *The horizon-T value functions satisfy $V_0^{\lambda}(m,d) = 0$ and*

$$V_T^{\lambda}(m,d) = \sup_{u \in \Delta^K} \left\{ \sum_{i=1}^{K} \left(u_i \Big(f(m,d,i) + \beta \mathbb{E}[V_{T-1}^{\lambda}(\Phi(m,d,i,Z))] \Big) \right) + \lambda \mathcal{H}(u) \right\}.$$

where Z is a random variable with the same distribution as Z_1.

In this subsection, we will use the functions defined in Section 3.4.2 and the dynamic programming in Theorem 3.4 to establish an approximation for the finite horizon value function with fixed λ.

$$V_T^{\lambda}(m,d) \approx \sum_{t=1}^{T} \beta^{T-t} \big(\mathrm{S}_{\max}^{\lambda} \circ \alpha_t^{\lambda} \big)(m,d). \tag{3.4.10}$$

Through the rest of this section, we will leave λ as a fixed constant, but will keep track of the order of error in terms of λ and $|d|$. For notational simplicity, we will omit superscript λ until the end of this section.

A heuristic description of our analysis

As discussed in the introduction of this chapter, when the arms does not depend on each other (i.e. when Gittins' theorem holds), the optimal strategy chooses based on an index strategy – the bandit chosen maximises the index α, which has the form

$$\alpha = \text{Exploitation gain} + \text{Learning premium}.$$

The exploitation gain is often simply the expected reward, while the learning premium corresponding to information gain.

Let see the learning premium term as an additional reward contributing to the future value function. In particular, we see α as an 'incremental reward' over a single step for each choice.

The term α is made up of the sum of an expected reward and the improvement of the value function in the future if the choice is taken. By recursively aggregating

these rewards, we obtain the value function. To outline this recursion, we will now give a heuristic version of our analysis.

The discounted value V_T: When there are t steps to go, we assume that *the incremental value* for each alternative is given by the vector $\alpha_t(m,d)$, introduced in (3.4.8). Hence, the optimal incremental reward, when there are t steps to go (equivalently, at the $(T-t)$th step), can be approximated by $\mathrm{S}_{\max}\big(\alpha_t(m,d)\big)$. After discounting, the value function V_T with horizon T can be estimated by

$$V_T(m,d) \approx \mathrm{S}_{\max}\big(\alpha_T(m,d)\big) + \beta\,\mathrm{S}_{\max}\big(\alpha_{T-1}(m,d)\big) + \cdots + \beta^{T-1}\,\mathrm{S}_{\max}\big(\alpha_1(m,d)\big).$$

This approximation will be formally obtained in Theorem 3.10.

The incremental value α_t: The ith component of $\alpha_t(m,d)$ can be interpreted as (an approximation to) the increase in the total expected value, when we choose option i with t steps to go. The term $\alpha_t^{(i)}(m,d)$ is therefore decomposed into two components:

(i) $f^{(i)}(m,d)$ is the reward obtained immediately after choosing option i.

(ii) $F_t^{(i)}(m,d) = \beta l_{t-1}^{(i)}(m,d) + ... + \beta^{t-1} l_1^{(i)}(m,d)$ can be seen as a learning term which describes how our decision affects the future value function. The function $l_s^{(i)}(m,d)$ describes the improvement (after choosing i) of the optimal single step value reward when there are s steps to go.

Now, let's suppose that the ith option is chosen. The underlying state (m,d) will change by

$$\Delta^{(i)}m = \mu^{(i)}(m,d) + \sigma^{(i)}(m,d)Z \qquad \text{and} \qquad \Delta^{(i)}d = b^{(i)}(m,d),$$

where $\mathbb{E}(Z) = 0$, $\mathrm{Var}(Z) = I_F$ and $\mathbb{E}|Z|^3 < \infty$. By Taylor's approximation, we show (in Corollary 3.9) that

$$\mathbb{E}\Big[\,\mathrm{S}_{\max}\big(\alpha_t(m+\Delta^{(i)}m, d+\Delta^{(i)}d)\big)\Big] \approx \mathrm{S}_{\max}\big(\alpha_t(m,d)\big) + l_t^{(i)}(m,d).$$

The appearance of ν, η and derivatives of f in $\mathcal{B}$, $\mathcal{M}$ and Σ arises because we consider the derivatives of $\mathrm{S}_{\max}\circ\alpha_t$ evaluated at (m,d) in Taylor's approximation (Lemma 3.3).

The function $l_t^{(i)}(m,d)$ can be seen as the change in the maximum single step value reward when there are t steps to go, if we choose the ith option (now). These terms are discounted and summed to give a single step incremental value for the

ith option. In particular, when there are $T+1$ steps remaining, it follows from the Dynamic Programming Principle that

$$V_{T+1}^{\lambda}(m,d) \approx \sup_{u\in\Delta_K} \Big\{ \sum_{i=1}^{K} u_i \Big(f(m,d,i) + \beta \sum_{t=1}^{T} \beta^{T-t} \big[\mathrm{S}_{\max}\big(\alpha_t(m,d)\big) + l_t^{(i)}(m,d) \big] \Big) + \lambda \mathcal{H}(u) \Big\}.$$

By rearranging the expression above, we can see that the single step value can be approximated by

$$\alpha_T^{(i)}(m,d) = f^{(i)}(m,d) + \beta l_{T-1}^{(i)}(m,d) + \cdots + \beta^{T-1} l_1^{(i)}(m,d).$$

Analysis of the learning value

We will now proceed to establish the heuristic approximation outlined above formally.

We have discussed that the incremental reward α_T can be decomposed into the instantaneous reward f and the learning term F_T. Each term l_t in F_T estimates the change in our expectations of future optimal rewards f, but does not take into account the additional change in future *learning* improvement arising from an improved decision today.

To justify this omission, we need to argue that when we make a decision, the contribution from the change in the learning term F_T is insignificant compared to the instantaneous cost f. The proof of each of these results (Lemma & Corollary 3.5 - 3.9) is simply a careful application of Taylor's expansion and can be found in Appendix A. We then use those results to prove Theorem 3.10.

We first recall our learning component, the function L, given in (3.4.7). We will describe how L_i changes when the state changes.

Lemma 3.5. *Let C be the constant upper bound in Assumption 3.2. Then there exist $P_2 \in \mathcal{P}_2^-$ and $q_2 \in \mathcal{P}_2^+$ such that for all $\lambda > 0$, $a, \Delta a \in \mathbb{R}^K, m, \Delta m \in \mathbb{R}^G, d, \Delta d \in \mathbb{R}^H$ with $|\Delta d| \le C|d|^2$,*

$$\big| L_i(a+\Delta a, m+\Delta m, d+\Delta d) - L_i(a,m,d) \big| \le P_2(\lambda) q_2(|d|) \big(|\Delta a| + |\Delta m| + |d| \big).$$

In order to ensure that the error in our approximation does not explode, we want to ensure that $|d|$ is sufficiently small.

Definition 3.5. Let $P_2(\lambda)$ and $q_2(h)$ be the upper bounds given in Lemma 3.5. We define the set

$$\mathcal{D}(\lambda) := \big\{ d \in \mathbb{R}_+^H : P_2(\lambda) q_2(|d|) \le 1-\beta \big\}.$$

Remark 3.9. One may replace $1-\beta$ in Definition 3.5 with any $\rho < (1-\beta)/\beta$. This can be seen as a trade-off between the coefficients of different errors in the approximation. We will restrict our consideration to $\rho = 1-\beta$ for notational simplicity.

By observing the recursive relation in (3.4.8) and apply Grönwall's inequality (Corollary A.2), we establish the order of the change in the learning term.

Lemma 3.6. *There exists $P_2 \in \mathcal{P}_2^-$, $q_2 \in \mathcal{P}_2^+$ such that for any $m, \Delta m \in \mathbb{R}^G, d, \Delta d \in \mathbb{R}^H$ with $d \in \mathcal{D}(\lambda)$ and $|\Delta d| \le C|d|^2$,*

$$\big|l_T(m+\Delta m, d+\Delta d) - l_T(m,d)\big| \le (1-\beta)^{-1} P_2(\lambda) q_2(|d|)\big(|\Delta m| + |d|\big).$$

Corollary 3.7. *There exists $P_2 \in \mathcal{P}_2^-$, $q_2 \in \mathcal{P}_2^+$ such that for $d \in \mathcal{D}(\lambda)$ and $|\Delta d| \le C|d|^2$,*

$$\big|F_T(m+\Delta m, d+\Delta d) - F_T(m,d)\big| \le (1-\beta)^{-2} P_2(\lambda) q_2(|d|)\big(|\Delta m| + |d|\big).$$

Analysis of the approximate value function and the approximate optimal control

By Assumptions 3.1 and 3.2, we notice that $|\mu(m,d)| \le C|d|^2$ and $|\sigma(m,d)| \le C|d|$. Hence, if we assume that $|d|$ is of order h (as $h \to 0$), we have that $\mathbb{E}|\Delta m|$ must be of order h. Therefore, Corollary 3.7 shows that the expected change in the learning term has order h^3. On the other hand, when the underlying state (m,d) changes, the instantaneous cost f still contributes a change of order h^2. Therefore, we omit the change arising due to the learning term F and only consider the change arising from the main instantaneous cost f.

We will now derive the relation between the single-step value reward α, the learning term F and how we can approximate the finite horizon value function via (3.4.10).

By considering a Taylor expansion (Lemma A.3 and Lemma A.4), one can show that:

Lemma 3.8. *There exists $P_2 \in \mathcal{P}_2^-$ and $q_3 \in \mathcal{P}_3^+$, such that for any $m, \Delta m \in \mathbb{R}^G, d, \Delta d \in \mathbb{R}^H$ with $d \in \mathcal{D}(\lambda)$ and $|\Delta d| \le C|d|^2$,*

$$\begin{aligned}
&\big(\mathrm{S}_{\max} \circ \alpha_T\big)(m+\Delta m, d+\Delta d) - \big(\mathrm{S}_{\max} \circ \alpha_T\big)(m,d) \\
&\quad = \Big\langle \partial_m \big(\mathrm{S}_{\max} \circ \alpha_T\big), \Delta m \Big\rangle + \Big\langle \partial_d \big(\mathrm{S}_{\max} \circ \alpha_T\big), \Delta d \Big\rangle + \frac{1}{2}\Big\langle \partial_m^2 \big(\mathrm{S}_{\max} \circ \alpha_T\big), \Delta m \Delta m^\top \Big\rangle + \Delta S_T
\end{aligned}$$

where $|\Delta S_T| \le (1-\beta)^{-2} P_2(\lambda)\big(|\Delta m|^3 + q_3(|d|)\big)$.

Corollary 3.9. *There exists $P_2 \in \mathcal{P}_2^-$ and $q_3 \in \mathcal{P}_3^+$, such that for any $m \in \mathbb{R}^G, d \in \mathbb{R}^H$ with $d \in \mathcal{D}(\lambda)$, writing $\Delta^{(i)}d = b^{(i)}(m,d)$ and $\Delta^{(i)}m = \mu^{(i)}(m,d) + \sigma^{(i)}(m,d)Z$, where $\mathbb{E}Z = 0$, $Var(Z) = I_F$ and $\mathbb{E}|Z|^3 < \infty$, we have*

$$\left|\mathbb{E}\left[\left(\mathrm{S}_{\max}\circ\alpha_T\right)(m+\Delta^{(i)}m, d+\Delta^{(i)}d)\right] - \left(\mathrm{S}_{\max}\circ\alpha_T\right)(m,d) - l_T^{(i)}(m,d)\right| \leq (1-\beta)^{-2}P_2(\lambda)q_3(|d|).$$

We now obtain an approximation for the change in the incremental value when our inverse precision d is sufficiently small. In order to use this approximation to approximate the value function (3.4.9), we need to make sure that d is sufficiently small through our process. Therefore, we will make the following assumption until the end of this section. For simplicity, we will specify the polynomials whose existence was given in Corollary 3.9 and Lemma 3.5 as a part of our assumption.

Assumption 3.4. *We assume the following: for a fixed $\lambda > 0$,*

(i) Suppose that we have $P_2 \in \mathcal{P}_2^-$, $q_2 \in \mathcal{P}_2^+$ and $q_3 \in \mathcal{P}_3^+$ such that inequality of Corollary 3.9 holds and

$$|L(a+\Delta a, m, d) - L(a, m, d)| \leq P_2(\lambda)q_2(|d|)|\Delta a|$$

for all $m \in \mathbb{R}^G$ and $d \in \mathcal{D}(\lambda) := \left\{d \in \mathbb{R}_+^H : P_2(\lambda)q_2(|d|) \leq 1-\beta\right\}$.

(ii) For the initial (information) state $D_0^U = d \in \mathbb{R}^H$, there exists a constant $h > 0$ such that

$$\sup_{U\in\mathcal{U}}\sup_{t\geq 0}|D_t^U| \leq h \qquad \text{and} \qquad P_2(\lambda)q_2(h) \leq (1-\beta).$$

Remark 3.10. The term h is a global bound on the (inverse) precision of our estimates. Here, we shall treat h as a small quantity in our analysis (i.e., we consider the setting where our estimates are precise). In the bandit problem, we usually have that $t \mapsto |D_t^U|$ is decreasing for all strategies $U \in \mathcal{U}$, so we can simply take $h = |d|$, provided $d \in \mathcal{D}(\lambda)$. Hence, we may interpret the following approximation as holding when we have a sufficiently large sample size.

Remark 3.11. It is worth pointing out that Assumption 3.4 was used for the convenience to state the results. This assumption fundamentally says that our initial precision d is small relative to the size of the smoothing parameter λ. We will no longer assume that this assumption holds when we allow λ to vary with the precision in section 3.5.

We can now use Assumption 3.4 together with the dynamic programming principle to give a formal approximation to $V_T(m,d)$.

Theorem 3.10. *Suppose that Assumption 3.4 holds, then*

$$\Big|V_T(m,d) - \sum_{t=1}^{T} \beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d)\Big| \le (1-\beta)^{-4}P_2(\lambda)q_3(h). \tag{3.4.11}$$

Proof. Define $\Delta V_T(m,d) = V_T(m,d) - \sum_{t=1}^{T}\beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d)$.

For each $T \ge 1$, we will choose (k_T) (possibly depending on h) sequentially such that for all m and d with $|d| \le h$, $|\Delta V_T(m,d)| \le k_T$. It is easy to see that $k_1 = 0$ via (3.4.4), (3.4.8) and (3.4.9).

Proceeding inductively, as in Corollary 3.9, write $\Delta^{(i)}d = b^{(i)}(m,d)$ and $\Delta^{(i)}m = \mu^{(i)}(m,d) + \sigma^{(i)}(m,d)Z$, where $\mathbb{E}Z = 0$, $\mathrm{Var}(Z) = I_F$ and $\mathbb{E}|Z|^3 < \infty$.

By Assumption 3.4, $|d + \Delta^{(i)}d| \le h$. Hence, by our inductive hypothesis, we have

$$\big|\Delta V_{T-1}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big| \le k_{T-1}.$$

Therefore, using the notation and result of Corollary 3.9,

$$\begin{aligned}
&\beta\mathbb{E}\big[V_{T-1}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big] \\
&= \sum_{t=1}^{T-1}\beta^{T-t}\mathbb{E}\big[\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big] + \beta\mathbb{E}\big[\Delta V_{T-1}^{(i)}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big] \\
&= \sum_{t=1}^{T-1}\Big(\beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d)\Big) + F_T^{(i)}(m,d) + \sum_{t=1}^{T-1}\beta^{T-t}\Delta_t^{(i)}(m,d) \\
&\qquad + \beta\mathbb{E}\big[\Delta V_{T-1}^{(i)}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big],
\end{aligned}$$

where, by Assumption 3.4 and the fact that $|d| \le h$,

$$|\Delta_t^{(i)}(m,d)| \le (1-\beta)^{-2}P_2(\lambda)q_3(|d|) \le (1-\beta)^{-2}P_2(\lambda)q_3(h).$$

Define

$$R_T^{(i)}(m,d) := \sum_{t=1}^{T-1}\beta^{T-t}\Delta_t^{(i)}(m,d) + \beta\mathbb{E}\big[\Delta V_{T-1}^{(i)}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)\big].$$

We then have $\big|R_T^{(i)}(m,d)\big| \le (1-\beta)^{-3}P_2(\lambda)q_3(h) + \beta k_{T-1}$.

By dynamic programming (Theorem 3.4),

$$\begin{aligned}
V_T(m,d) &= \sup_{u\in\Delta^K}\left\{\sum_{i=1}^{K}u_i\Big(f^{(i)}(m,d) + \beta\mathbb{E}[V_{T-1}(m + \Delta^{(i)}m, d + \Delta^{(i)}d)]\Big) + \lambda\mathcal{H}(u)\right\} \\
&= \sup_{u\in\Delta^K}\left\{\sum_{i=1}^{K}u_i\Big(\sum_{t=1}^{T-1}\beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d) + f^{(i)}(m,d) + F_T^{(i)}(m,d) + R_T^{(i)}(m,d)\Big) + \lambda\mathcal{H}(u)\right\} \\
&= \sum_{t=1}^{T-1}\beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d) + \sup_{u\in\Delta^K}\left\{\sum_{i=1}^{K}u_i\Big(\alpha_T^{(i)}(m,d) + R_T^{(i)}(m,d)\Big) + \lambda\mathcal{H}(u)\right\}.
\end{aligned}$$

Hence, it follows from the definition of $\Delta V_T(m,d)$ that

$$\sup_{u\in\Delta^K}\left\{\sum_{i=1}^{K}u_i\Big(\alpha_T^{(i)}(m,d)+R_T^{(i)}(m,d)\Big)+\lambda\mathcal{H}(u)\right\}=\big(\mathrm{S}_{\max}\circ\alpha_T\big)(m,d)+\Delta V_T(m,d),$$

where $|\Delta V_T(m,d)|\le\sup_{|d|\le h,m}|R_T(m,d)|\le(1-\beta)^{-3}P_2(\lambda)q_3(h)+\beta k_{T-1}$.

Therefore, given k_{T-1}, we can choose $k_T=(1-\beta)^{-3}P_2(\lambda)q_3(h)+\beta k_{T-1},$. As $k_1=0$, one can show by induction that (k_t) is increasing. Therefore, we obtain $k_T\le k_\infty=(1-\beta)^{-4}P_2(\lambda)q_3(h)$. □

3.4.4 From finite horizon to infinite horizon

In the previous subsection, we analysed how the value function V_T and the incremental reward α_T can be expressed through a recursive formula. In the recursive equation (3.4.8), a heavier weight is placed on $L(\alpha_t(m,d),m,d)$ when t is large. This suggests that the incremental reward α_T should converge to some fixed point, which we now describe.

Recall the function $L:\mathbb{R}^K\times\mathbb{R}^G\times\mathbb{R}^H\to\mathbb{R}^K$ given in (3.4.7) and consider $(m,d)\in\mathbb{R}^G\times\mathbb{R}^H$ such that Assumption 3.4 holds.

Define the map $\mathcal{T}:\mathbb{R}^K\to\mathbb{R}^K$ by $\mathcal{T}(a)=f(m,d)+\big(\frac{\beta}{1-\beta}\big)L\big(a,m,d\big)$.

By Assumption 3.4,

$$|L\big(a+\Delta a,m,d\big)-L\big(a,m,d\big)|\le P_2(\lambda)q_2(h)|\Delta a|\le(1-\beta)|\Delta a|.$$

Hence, $\mathcal{T}$ must be a contraction. Therefore, the fixed point exists and is unique.

Proposition 3.11. *For any $(m,d)\in\mathbb{R}^G\times\mathbb{R}^H$ such that Assumption 3.4 holds, there exists a unique $a\in\mathbb{R}^K$ such that $a=f(m,d)+\big(\frac{\beta}{1-\beta}\big)L\big(a,m,d\big)$.*

Definition 3.6. Suppose Assumption 3.4 holds. We define a function $\alpha:\mathbb{R}^G\times\mathbb{R}^H\to\mathbb{R}^K$ such that $\alpha(m,d)$ is the unique fixed point of

$$\alpha(m,d)=f(m,d)+\Big(\frac{\beta}{1-\beta}\Big)L\big(\alpha(m,d),m,d\big).\tag{3.4.12}$$

Theorem 3.12. *Suppose Assumption 3.4 holds. Then $(\alpha_T(m,d))$ given in (3.4.8) satisfies*

$$\sup_{|d|\le h,m}\Big|\alpha_T(m,d)-\alpha(m,d)\Big|\to 0\qquad as\qquad T\to\infty.$$

Proof. It suffices to show that $(\alpha_T(m,d))$ converges. Then the recursive formula (3.4.8) for α_T simplifies to (3.4.12); giving the desired equality.

We will show that (α_T) is a Cauchy sequence under the stated supremum norm. By Assumption 3.4, writing $\rho = 1-\beta$,

$$\big|L(a_1,m,d) - L(a_2,m,d)\big| \le P_2(\lambda)q_2(h)|a_1-a_2| \le \rho|a_1-a_2|.$$

Moreover, by Assumptions 3.2 and 3.3, there exists a polynomial $P_1 \in \mathcal{P}_1^-$ which does not depend on (a,m,d) such that for all $a \in \mathbb{R}^K$ and $(m,d) \in \mathbb{R}^{G+H}$, $|L(a,m,d)| \le P_1(\lambda)$.

Fix $s > 0$. Define $k_T = \sup_{|d|\le h,m} |\alpha_{T+s}(m,d) - \alpha_T(m,d)|$. Then

$$\begin{aligned}
k_T &= \sup_{|d|\le h,m} \left| \sum_{t=1}^{s+T-1} \beta^t L\big(\alpha_{T+s-t}(m,d),m,d\big) - \sum_{t=1}^{T-1} \beta^t L\big(\alpha_{T-t}(m,d),m,d\big)\right| \\
&\le \sum_{t=1}^{T-1} \beta^t \left(\sup_{|d|\le h,m} \left| L\big(\alpha_{T+s-t}(m,d),m,d\big) - L\big(\alpha_{T-t}(m,d),m,d\big)\right|\right) \\
&\quad + \sum_{t=T}^{T+s} \beta^t \left(\sup_{|d|\le h,m} \left| L\big(\alpha_{T+s-t}(m,d),m,d\big)\right|\right) \\
&\le \sum_{t=1}^{T-1} \beta^t \rho \left(\sup_{|d|\le h,m} \left|\alpha_{T+s-t}(m,d) - \alpha_{T-t}(m,d)\right|\right) + \beta^T P_1(\lambda)\Big(\frac{1}{1-\beta}\Big) \\
&= \rho \sum_{t=1}^{T-1} \beta^t k_{T-t} + \beta^T P_1(\lambda)\Big(\frac{1}{1-\beta}\Big).
\end{aligned}$$

Write $y_t = \beta^{-t} k_t$ for $t \ge 0$. Then

$$y_T \le P_1(\lambda)\Big(\frac{1}{1-\beta}\Big) + \rho \sum_{t=1}^{T-1} y_{T-t} = P_1(\lambda)\Big(\frac{1}{1-\beta}\Big) + \rho \sum_{t=1}^{T-1} y_t.$$

By Grönwall's inequality (Lemma A.1) with $c_t = P_1(\lambda)/(1-\beta)$,

$$y_T \le P_1(\lambda)\Big(\frac{1}{1-\beta}\Big)\Big(1 + \Big(\frac{\rho}{1+\rho}\Big)\sum_{t=1}^{T-1}(1+\rho)^t\Big) = P_1(\lambda)\Big(\frac{1}{1-\beta}\Big)(1+\rho)^{T-1}.$$

Since $0 < \beta(1+\rho) < \beta(1+(1-\beta)/\beta) = 1$, it follows that for all $s > 0$, as $T \to \infty$

$$\sup_{m,d\in\mathbb{R}^N,|d|\le h} \big|\alpha_{T+s}(m,d) - \alpha_T(m,d)\big| = k_T = \beta^T y_T \le P_1(\lambda)\Big(\frac{1}{1-\beta}\Big)(1+\rho)^{-1}\big(\beta(1+\rho)\big)^T \to 0.$$

Hence, α_T is a Cauchy sequence and thus converges. □

So far, we have shown that α_T converges and the horizon-T value function V_T can be approximated in terms of α_T. Notice that the expression for our approximation

of V_T (3.4.11) is an exponential weighted average of functions of α_t. As α_t converges, we also expect our approximation for V_T to converge.

We now obtain an approximation for the infinite horizon problem (3.4.2). To prove this result, we use the boundedness of f and $\mathcal{H}$ to show that $V_T \to V_\infty$ as $T \to \infty$ (Lemma A.5). We then use Tauberian theorem (Proposition A.6) to obtain convergence of our approximation $\sum_t \beta^{T-t}(\mathrm{S}_{\max} \circ \alpha_t) \to \big(\frac{1}{1-\beta}\big)(\mathrm{S}_{\max} \circ \alpha)$ as $T \to \infty$. We can now obtain

Theorem 3.13. *Let α be the function defined in* (3.4.12)*. If Assumption 3.4 holds, then*

$$\sup_{|d| \le h, m} \Big| V_\infty(m,d) - \Big(\frac{1}{1-\beta}\Big)\big(\mathrm{S}_{\max} \circ \alpha\big)(m,d) \Big| \le (1-\beta)^{-4} P_2(\lambda) q_3(h).$$

3.5 Collapse in the entropy term

In the previous section, we discussed an approximate solution to (3.4.2) and (3.4.9). To obtain an approximation, we introduced an entropy as an additional reward for our decision, which smoothes our value function. However, the objective of our problem is to solve the optimisation problem without the entropy,

$$V(m,d) \coloneqq \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty} \beta^t \Big(f\big(M_t^U, D_t^U, A(U_{t+1}, \zeta_{t+1})\big)\Big)\Big]. \tag{3.5.1}$$

where the drift and the volatility of the underlying process $X = (M, D)$ satisfy Assumption 3.1 and 3.2.

As $\mathcal{H}$ is assumed bounded, introducing the entropy with a fixed scaling factor λ will give an error in the value function of order $\mathcal{O}(\lambda)$ (via the same argument as in Lemma A.3). Hence, in order to propose a good decision, we may wish to let λ collapse through time, i.e., we may consider λ as a function of (m,d). By doing this, we can ensure that we approach the optimum of (3.5.1).

In Theorem 3.13, as $P_2 \in \mathcal{P}_2^-$ and $q_3 \in \mathcal{P}_3^+$, the error of our approximation is of order $\mathcal{O}(h^3/\lambda^2)$, where h is an upper bound on the process D_t started from d. Moreover, we can also observe that the difference between (3.2.2) and (3.4.2) is of order $\mathcal{O}(\lambda)$. Hence, in order to have an error of the approximation converging to zero, we choose λ to be of order $\mathcal{O}(h^s)$ where $s \in (0, 3/2)$, with $s = 1$ giving a natural trade-off between our error terms. However, Assumption 3.4 requires $P_2(\lambda)q_2(h) < (1-\beta)$. This means that, when $t \mapsto |D_t|$ is assumed decreasing (so $h = |d|$), the minimal order of $\lambda(d)$ must be $\mathcal{O}(h^s)$ where $s > 1$, i.e. we are restricted to consider $s \in (1, 3/2)$.

In this section, we will discuss the choice of λ, allowing λ to decay through time in the same manner as describing above. We will only focus on the bandit problem with an unknown parameter with no contextual element for convenience in our analysis. Through the remainder of this chapter, we will make the following assumption (where the contextual bandit often fails to satisfy).

Assumption 3.5. *For any $d \in \mathbb{R}_+^H$, the process $t \mapsto |D_t^U|$ with $D_0^U = d$ is non-increasing for all $U \in \mathcal{U}$.*

To allow variety, on the choice of λ and allowing λ to vary through time, we consider the following definition.

Definition 3.7. Let $P_2(\lambda)$ and $q_2(h)$ be the bounds such that the inequality in Lemma 3.5 holds. We say a continuous function $\lambda : \mathbb{R}^G \times \mathbb{R}_+^H \to \mathbb{R}_+$ is *consistent* if $\sup_m \lambda(m,d) \downarrow 0$ as $d \downarrow 0$ and

$$0 < h^* := \sup\Big\{h : \sup_{m,d:|d|\le h} P_2(\lambda(m,d))q_2(|d|) \le (1-\beta)\Big\}.$$

We call h^* *the consistency radius for* λ.

Example 3.2. For $k : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, the function $\lambda(m,d) = \Big(\sum_{i=1}^K k(m_i,d_i)d_i^{2+\epsilon}\Big)^{1/2}$ is consistent when $\epsilon > 0$ and k is bounded above and away from zero.

As discussed in Remark 3.8, in order to solve (3.5.1), our approximate optimal decision with $\lambda = \lambda(m,d)$ is given by $\nu^{\lambda(m,d)}\big(\alpha^{\lambda(m,d)}(m,d)\big)$, where $\nu^\lambda(a) = \arg\max_{u\in\Delta^K}\Big(\sum_{i=1}^K u_i a_i + \lambda\mathcal{H}(u)\Big) = \partial_y S\Big|_{y=a/\lambda}$ and $\alpha^{\lambda(m,d)}(m,d)$ satisfies

$$\alpha = f(m,d) + \Big(\frac{\beta}{1-\beta}\Big)L^{\lambda(m,d)}\big(\alpha,m,d\big). \tag{3.5.2}$$

For the rest of this section, we will show that this feedback control is a good choice to optimise (3.5.1). In particular, we will show that, if $\lambda(m,d)$ is consistent, we have:

(i) When d is sufficiently small, the feedback control $\nu^{\lambda(m,d)}\Big(\alpha^{\lambda(m,d)}(m,d)\Big)$ yields an error, from the optimal value (3.5.1), converging to 0 as $d \to 0$.

(ii) If we follow the feedback control $\nu^{\lambda(m,d)}\Big(\alpha^{\lambda(m,d)}(m,d)\Big)$, then the corresponding information process D_t converges to 0 almost surely. In particular, $\sup_m \lambda(m,d) \to 0$ as $d \to 0$, and we asymptotically achieve a solution to (3.5.1).

3.5.1 Near Optimality of the feedback control

When $|d|$ is sufficiently small, we can see that (3.5.2) has a solution. We can thus define our corresponding feedback control to be

$$U^{\lambda} = \Big(\nu^{\lambda(M_t,D_t)}\big(\alpha^{\lambda(M_t,D_t)}(M_t,D_t)\big)\Big)_{t\geq 0}.$$

We also define the value function corresponding to a strategy U to be

$$V^{U}(m,d) := \mathbb{E}_{m,d}\Big[\sum_{t=0}^{\infty}\beta^t f\big(M_t^U, D_t^U, A_{t+1}\big)\Big].$$

To simplify our discussion, we also define the one-step Q-function for our optimisation problem.

For a fixed $u \in \Delta^K$, define

$$Q(u,m,d) := \sum_{i=1}^{K}\Big(u_i\Big(f(m,d,i) + \beta\mathbb{E}[V(\Phi(m,d,i,Z))]\Big)\Big).$$

By considering various approximation scheme to Q and V, we can show that using the feedback control $\nu^{\lambda(m,d)}\Big(\alpha^{\lambda(m,d)}(m,d)\Big)$ once and then switching to the optimal strategy results a relatively small error to the optimal value. We then use this to quantify the error when we only follow the proposed control.

Theorem 3.14. *Let $\lambda(m,d)$ be consistent with consistency radius h^* and let $\alpha^{\lambda(m,d)}(m,d)$ be the solution to (3.4.12) for $|d| \leq h^*$. There exists a constant C (possibly depending on β but not m,d) and a function $q_1 \in \mathcal{P}_1^+$ such that, for all $|d| \leq h^*$,*

$$Q\Big(\nu^{\lambda(m,d)}\big(\alpha^{\lambda(m,d)}(m,d)\big),m,d\Big) + C\big(\lambda(m,d) + q_1(|d|)\big) \geq V(m,d). \tag{3.5.3}$$

Corollary 3.15. *Define the function .*

Let $\lambda(m,d)$ be consistent with consistency radius h^* and let $\alpha^{\lambda(m,d)}(m,d)$ be the solution to (3.4.12) for $|d| \leq h^*$. There exists a constant C (possibly depending on β but not m,d) and a function $q_1 \in \mathcal{P}_1^+$ such that, for all $|d| \leq h^*$,

$$V^{U^{\lambda}}(m,d) + C\big(\sup_m \lambda(m,d) + q_1(|d|)\big) \geq V(m,d). \tag{3.5.4}$$

3.5.2 Convergence of the information

To finish our theoretical discussion, we will show that as $t \to \infty$, we will fully explore our system. In particular, we will show that the inverse precision d converges to 0 almost surely as $t \to \infty$.

In order to ensure that it is possible to fully explore the system, we assume that if we keep playing any chosen arm, we will know more about the chosen arm and will have perfect knowledge after infinitely many plays. This leads us to the following definition.

Definition 3.8. We say a family of drifts $\left(b^{(i)}\right)_{i\in[K]}$ is *asymptotically well informed* if, for every $i \in [K]$, $b^{(i)}$ satisfies: for any sequence of (m_t) taking values in $\mathbb{R}^G$, if we have a sequence $(d_t) \subseteq \mathbb{R}^H_+$ satisfying $d_{t+1} \in [0, d_t + b^{(i)}(m_t, d_t)]$ where the interval is interpreted component-wise, it follows that $d_t^{(i)} \to 0$ as $t \to \infty$.

Remark 3.12. It is easy to show that the drifts in (3.3.3) and (3.3.6) derived for a multi-arm bandit and a correlated bandit are asymptotically well informed.

Remark 3.13. In Proposition 3.11, we show that the solution $\alpha^\lambda(m, d)$ exists when $d \in \mathcal{D}(\lambda)$. In particular, we have shown that the solution $\alpha^{\lambda(m,d)}(m, d)$ to (3.4.12) exists and is unique when $|d| \le h^*$ where h^* is a consistency radius of λ. In order to ensure that our strategy U^λ is well-defined for all $(m, d) \in \mathbb{R}^G \times \mathbb{R}^H$. We may define $\alpha^{\lambda(m,d)}(m, d)$ to be a root of (3.4.12) when a solution exists and define

$$\alpha^{\lambda(m,d)}(m,d) = f(m,d) + \Big(\frac{\beta}{1-\beta}\Big) L^{\lambda(m,d)}\big(f(m,d), m, d\big)$$

otherwise.

Theorem 3.16. *Let S be a smooth max approximator. Suppose that for any compact set $K \subseteq \mathbb{R}^K$, there exists a non-empty open ball, $B(r)$, such that $B(r) \cap S'(K) = \phi$.*

Suppose further that the drifts $\big(b^{(i)}(m, d)\big)$ are asymptotically well informed, $\lambda(m, d)$ is consistent and $d \to \inf_m \lambda(m, d)$ is continuous and strictly positive.

Let M_t and D_t be a process representing the underlying state corresponding to the feedback control

$$U = \Big(\nu^{\lambda(M_t,D_t)}\big(\alpha^{\lambda(M_t,D_t)}(M_t, D_t)\big)\Big)_{t\in\mathbb{N}_0},$$

where $\alpha^{\lambda(m,d)}(m, d)$ is chosen as discussed in Remark 3.13.

Then for any $D_0 \in \mathbb{R}^H$ with $D_0 \ge 0$, $|D_t| \to 0$ a.s.

Proof. We will show that on the event that $|D_t| \not\to 0$, every bandit will be played infinitely many times, almost surely.

Fix $i \in \mathcal{A}$ and $\epsilon > 0$. Define $A_\epsilon^{(i)} := \{D_t^{(i)} > \epsilon \text{ for all } t \ge 0\}$. We will first show that $A_\epsilon^{(i)}$ is a null-set.

By a similar argument as in Lemma 3.5, we can find $P_1 \in \mathcal{P}_1^-$ such that L_j^λ can be uniformly bounded by some $P_1(\lambda)$ for all $j \in \mathcal{A}$. Now, as $d \mapsto \inf_m \lambda(m, d)$ is

continuous and strictly positive, there exists $\lambda_0 > 0$ such that $\lambda(m,d) > \lambda_0$ for all $m \in \mathbb{R}^G$ and $d \in \{d \in \mathbb{R}_+^P : \epsilon \le |d| \le |D_0|\}$. Hence, $|L_j^{\lambda(m,d)}(a,m,d)| \le P_1(\lambda(m,d)) \le P_1(\lambda_0)$ for all $j \in \mathcal{A}$. In particular, $\sup_a L^\lambda(a, M_t, D_t)$ is uniformly bounded on $A_\epsilon^{(i)}$.

By our choice of α, $\alpha^{\lambda(m,d)}(m,d) = f(m,d) + \left(\frac{\beta}{1-\beta}\right) L^{\lambda(m,d)}\big(g(m,d),m,d\big)$, where either $g(m,d) = f(m,d)$ or $g(m,d) = \alpha^{\lambda(m,d)}(m,d)$. As f is bounded by Assumption 3.3, $\alpha^{\lambda(M_t,D_t)}(M_t,D_t)$ must take values in a compact set on $A_\epsilon^{(i)}$.

Moreover, as $\lambda(M_t,D_t) > \lambda_0$ on $A_\epsilon^{(i)}$, $\alpha^{\lambda(M_t,D_t)}(M_t,D_t)/\lambda(M_t,D_t)$ must take values in some compact set K. By our assumption on S, we can find $r > 0$ such that $B(r) \notin S'(K)$. In particular, for any $a \in K$, we know $S'(a_i) > r > 0$.

Recall that $\nu^\lambda(a) = S'(a/\lambda)$. Hence, on the event $A_\epsilon^{(i)}$, $\nu_i^{\lambda(M_t,D_t)}\Big(\alpha^{\lambda(M_t,D_t)}(M_t,D_t)\Big) > r > 0$.

As $\nu_i^{\lambda(M_t,D_t)}\Big(\alpha^{\lambda(M_t,D_t)}(M_t,D_t)\Big)$ is the probability to choose the ith option at time t, conditional on $A_\epsilon^{(i)}$, the ith option must be chosen infinitely often. By well-informedness of $b^{(i)}(m,d)$ (Definition 3.8), it follows that $D_t^{(i)} \to 0$ as $t \to \infty$ on $A_\epsilon^{(i)}$ (almost surely). Hence, $A_\epsilon^{(i)}$ must be a null set for all $\epsilon > 0$ otherwise we have a contradiction.

By considering the event $A = \bigcup_{i\in\mathcal{A}} \bigcup_{\epsilon\in\mathbb{Q}_+} A_\epsilon^{(i)}$, the result follows. □

Remark 3.14. The assumption of the existence of $B(r)$ can be interpreted as follows: if the value difference between any two options is bounded, then all options must be considered, with a nonzero probability of selection.

Remark 3.15. In Theorem 3.14 and Corollary 3.15, we require our inverse precision d to be reasonably small to guarantee a low error in its decision. In contrast, Theorem 3.16 shows that no matter what the initial level of precision, we will eventually (with probability 1) reach a point where d is sufficiently small to obtain inequalities in Theorem 3.14 and Corollary 3.15. Nevertheless, if we start with a low level of initial information (i.e., $|d|$ is large), the probability of entering the regime where Theorem 3.14 holds, within a reasonable time, could be low. Therefore, we may want to start applying our algorithm when we have a reasonable sample size from each arm to ensure that we will achieve (3.5.3) sooner.

Chapter 4

Performance of the Asymptotic Randomised Control (ARC): Numerical simulation

In Chapter 3, we consider the multi-armed bandit problem as a relaxed control problem and obtain its asymptotic approximation. This approximation gives a simple strategy to tackle a general class of multi-armed bandit problems by solving a fixed point equation. We will call the resulting strategy the 'Asymptotic Randomised Control (ARC) algorithm.'

We have already discussed the theoretical properties of the ARC algorithm in the earlier section. We have shown that this algorithm gives a near-optimal total discounted payoff. In this chapter, we will see the ARC algorithm from a practical perspective. We will run a simulation in a simple bandit setting and compare the ARC algorithm with other algorithms discussed in Chapter 2.

For the convenience of the reader, we will first summarise our results from Chapter 3 in the form of an algorithm for a multi-armed bandit problem in Section 4.1. In Section 4.2, we consider a simple bandit problem, find an optimal solution explicitly, and compare our approximation to the computed optimal solution. Finally, we compare our algorithm with other algorithms for the bandit in Section 4.3.

This chapter is based on the paper [34].

4.1 Asymptotic Randomised Control (ARC) algorithm

We will summarise how our approximation of the control problem yields an explicit algorithm for a general class of multi-armed bandits. This algorithm is fully justified

as an asymptotic approximation when the dynamics of the posterior parameter and the chosen decision parameters λ, S, and β satisfy all assumptions, as demonstrated in the previous chapter.

Let $R^{(i)}(Y^{(i)})$ be the reward when the ith option is chosen and we observe the outcome $Y^{(i)}$. We suppose $Y^{(i)}$ is sampled from a distribution $\mu_\Theta^{(i)}$,where Θ is unknown. Suppose $\mu_\Theta^{(i)}$ governs a prior-posterior conjugate pair and Θ has a prior/posterior described by parameters $(m, d) \in \mathbb{R}^G \times \mathbb{R}_+^H$. We think of m as a location parameter for Θ, while d indicates the degree of uncertainty we have in our estimates, which should be small and converge to zero as the number of observations tends to infinity.

We define the conditional expectation $f(m, d, i) \coloneqq \mathbb{E}_{m,d}\Big(R^{(i)}(Y^{(i)})\Big)$ where the expectation is taken with (m, d) as a prior of Θ.

Suppose that, after taking action i and making the corresponding observations, the posterior distribution of Θ is described by the parameters $(M^{(i)}, D^{(i)})$. This yields the parameter dynamics:

$$\begin{aligned}
\mu^{(i)}(m, d) &\coloneqq \mathbb{E}_{m,d}\big(M^{(i)}\big) - m, \\
\big(\sigma\sigma^\top\big)^{(i)}(m, d) &\coloneqq \mathrm{Var}_{m,d}\big(M^{(i)}\big), \\
b^{(i)}(m, d) &\coloneqq \mathbb{E}_{m,d}\big(D^{(i)}\big) - d.
\end{aligned}$$

Next, we choose a smooth max approximator S satisfying Definition 3.3 and Theorem 3.2 (a simple choice for S is $S(a) = \log\big(\sum_i \exp(a_i)\big)$, as in Example 3.1). We also choose a function $\lambda : \mathbb{R}^G \times \mathbb{R}_+^H \to \mathbb{R}_+$ which satisfies Theorem 3.16.

We propose to consider

$$\lambda = \lambda_\rho(m, d) \coloneqq \rho\Big(\sum_{i=1}^K S'(f(m, d))_i d_i^2\Big)^{1/2}, \tag{4.1.1}$$

where the parameter ρ is chosen by the user; this does not quite satisfy Definition 3.7, but we can ensure the required assumptions by adding a small perturbation on the power of d_i which has a negligible effect. We thus encourage this choice of λ for convenience.

Remark 4.1. In Section 3.3, we introduce the ith component of d as a precision corresponding to the ith arm.The bad arms are expected to be played infrequently. Thus, λ may decay too slow, resulting in too little exploitation. To avoid this effect, we may add some weight to the precision when evaluating λ. This weight is given in terms of $S'(f)$ in (4.1.1) which puts more weight toward a good arm.

Define $L^\lambda : \mathbb{R}^K \times \mathbb{R}^G \times \mathbb{R}^H \to \mathbb{R}^K$, by its components

$$L_i^\lambda(a,m,d) := \langle \mathcal{B}^\lambda(a,m,d); b^{(i)}(m,d)\rangle + \langle \mathcal{M}^\lambda(a,m,d); \mu^{(i)}(m,d)\rangle + \frac{1}{2}\langle \Sigma^\lambda(a,m,d); \left(\sigma\sigma^\top\right)^{(i)}(m,d)\rangle,$$

where $\mathcal{B}^\lambda$, $\mathcal{M}^\lambda$ and Σ^λ are functions of λ and derivatives of f and S, which are defined in Definition 3.4 and in Lemma 3.3. This function L^λ gives the key quantity for the ARC algorithm, which we can now define.

Algorithm 1 ARC(ρ, β, m, d)

1: Define $\lambda = \lambda_\rho(m,d)$ as in (4.1.1)
2: Numerically, find a satisfying $a = f(m,d) + \left(\frac{\beta}{1-\beta}\right)L^\lambda\left(a,m,d\right)$
If such an a does not exist, define $a = f(m,d) + \left(\frac{\beta}{1-\beta}\right)L^\lambda\left(f(m,d),m,d\right)$.
3: Sample $i \sim \text{Random}([K], S'(a/\lambda))$
4: **return** i

Remark 4.2. If more than one such a exists, the choice of a may be determined by the numerical algorithm used. In our implementation, we use the standard `scipy.optimize.root` solver in python, which uses a modified Powell conjugate gradient method, with initialization at the point $a = f(m,d) + \left(\frac{\beta}{1-\beta}\right)L^\lambda\left(f(m,d),m,d\right)$.

We can then apply this algorithm to a multi-armed bandit problem.

Algorithm 2 MultiArmedBandit(T, m, d, Algorithm)

1: **for** $t = 1, ..., T$ **do**
2: Choose i = Algorithm(m,d)
3: Observe $Y^{(i)}$ and collect reward $R^{(i)}(Y^{(i)})$
4: Update parameter m and d
5: **end for**

4.2 Comparisons with optimal solution

To illustrate that the ARC algorithm gives a reasonable answer, we will compare our estimated value function and its corresponding control (the ARC strategy) to the exact value function that can be computed in a simple setting.

Suppose that we have two arms. One arm always give us a reward of 1/2 and another give a reward $\Phi(Y)$ where Φ is the cumulative density function of $N(0,1)$, $Y \sim N(\Theta, 1)$ and Θ is unknown. We will write (m,d) for the (prior/posterior) mean and variance of Θ.

For this simple case, we can compute an accurate value function

$$V(m,d) = \sup_{U \in \mathcal{U}} \mathbb{E}_{m,d}\left[\sum_{t=0}^{\infty} \beta^t \left(R^{(A_{t+1})}\left(Y_{t+1}^{(A_{t+1})}\right)\right)\right],$$

where we set $\beta = 0.99$, using backward induction (together with Monte–Carlo simulation). In particular, we start our iteration with $V^0(m,d) = 0$. Writing $f(m,d)$ for the vector of expected rewards given the state (m,d), we iteratively use Monte-Carlo simulation to compute

$$V^{n+1}(m,d) = \max_i \left(f_i(m,d) + \beta \mathbb{E}(V^n(M_1,D_1)|M_0 = m, D_0 = d, A_1 = i)\right)$$

on a grid of (m,d) and use interpolation to approximate the function V^{n+1}. We then repeat the procedure until the function converges.

We can also use our approximation (3.4.3) to compute an estimated value

$$V_\rho(m,d) := \left(\frac{1}{1-\beta}\right)\left(S_{\max}^{\lambda_\rho(m,d)} \circ \alpha^{\lambda_\rho(m,d)}\right)(m,d)$$

where we here consider $\lambda_\rho(m,d)$ as in (4.1.1) with $S(a) = \log\left(\sum_{i=1}^K \exp(a_i)\right)$ (In this case, $K = 2$).

The corresponding randomised decision, given an estimated parameter ρ, can be calculated by

$$U_\rho(m,d) = \nu^{\lambda_\rho(m,d)}\left(\alpha^{\lambda_\rho(m,d)}(m,d)\right).$$

In Figure 4.1, we can see that when ρ takes a reasonable value, our estimated value function is close to the exact value.

Remark 4.3. Our approximate value function appears discontinuous when the number of observations (n) is low. (i.e. the inverse precision ($d = 1/n$) is large.) For such values of d (and $\lambda_\rho(m,d)$), the uniqueness result of Proposition 3.11 does not hold. However, one can show explicitly that, in this case, at least one solution to (3.5.2) exists but may not be unique for some value of d. This non-uniqueness of the solution results in the jump in our approximated value function.

In Figure 4.2, in the top-left plot, we illustrate the optimal probability ($U^*(m,d)$) of playing the unknown arm. The remaining plots show the difference in our approximation ($U^*(m,d) - U_\rho(m,d)$), for various values of ρ. We can see that our approximate decision is more likely to stop exploring (i.e., start playing a known option) too early. Nonetheless, we can see that when the value of ρ increases, the difference becomes smaller. Moreover, when the precision is high (i.e., d is low), our algorithms are close to the optimal decision.

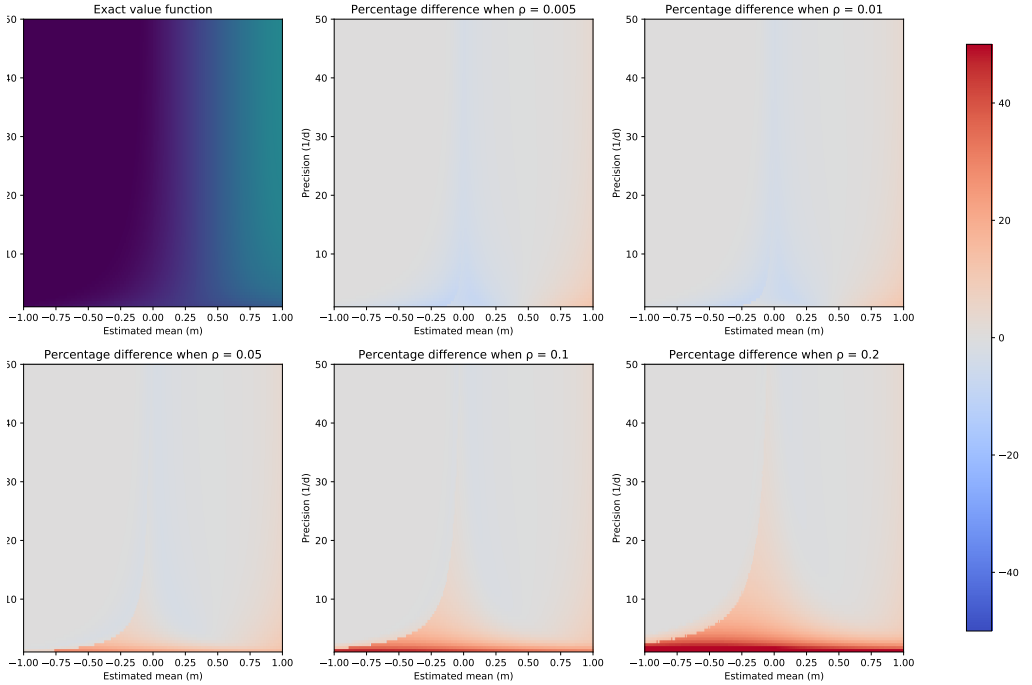

Figure 4.1: Value function and the percentage difference for ARC approximation

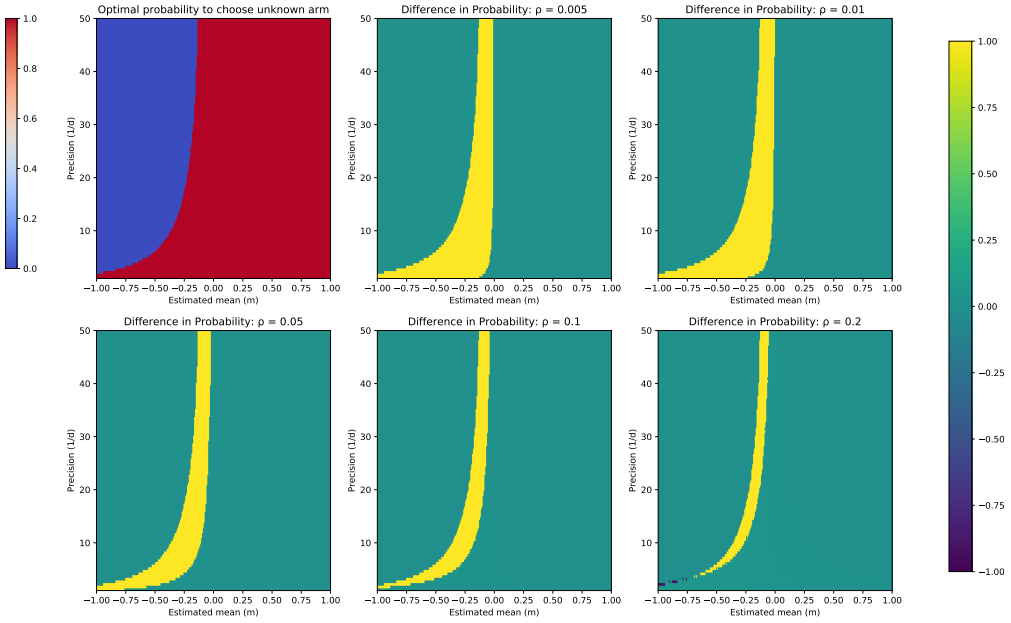

Figure 4.2: The optimal probability to choose the unknown arm and the difference between the optimal probability and the ARC probabilistic decision.

Remark 4.4. Our asymptotic expansion gives $\tilde{V}_\rho(m,d) := \left(\frac{1}{1-\beta}\right) S_{\max}^{\lambda_\rho(m,d)} \circ \alpha^{\lambda_\rho(m,d)}(m,d)$ as an approximation to $V(m,d)$, in the regime where d is small. Instead of directly using the corresponding asymptotic policy, we may combine this approximation with a single step of the Bellman equation, in a similar way to the Knowledge Gradient Algorithm (Ryzhov et al. [100]).

In particular, we can estimate the optimal policy

$$U_\rho(m,d) = \arg\max_{i \in \mathcal{A}} \Big(f(m,d,i) + \beta \mathbb{E}\big(\tilde{V}_\rho(M_1, D_1) \big| M_0 = m, D_0 = d, A_1 = i\big)\Big).$$

A challenge in implementing this approach is the large number of fixed-point problems that must be solved. In the case considered earlier, we can instead compute $\tilde{V}_\rho(m,d)$ on some grid of (m,d) and use an interpolation to evaluate the function $\tilde{V}_\rho$. We can then use a Monte-Carlo simulation as previous to compute the term in the expectation.

In general, when the dimension of parameters is large, one can use regression (or other function approximation techniques) to give an approximation of $\tilde{V}_\rho$ based on a limited number of fixed-point calculations.

We will leave this computational path of research, and evaluation of the resulting performance of this mixed asymptotic–Monte-Carlo algorithm, for further study.

4.3 Performance relative to other bandit algorithms

In this section, we will consider the performance of our algorithm for the multi-armed bandit problem discussed in Section 3.3.

For simplicity of our simulation, we consider $R^{(i)}(y) := \Phi(y) - r_i$ where $r_i \in \mathbb{R}$ and Φ is the cumulative density of $N(0,1)$. We may see r_i as a fixed cost and see $\Phi(y)$ as a reward to play the ith arm.

We will consider $S(a) = \log(\sum_{i=1}^K \exp(a_i))$ and consider $\lambda_\rho(m,d)$ as in (4.1.1) to implement the ARC algorithm. We will also set $\beta = 1 - 1/T$ where T is the horizon of the problem. This choice is inspired by the Knowledge Gradient algorithm (Ryzhov et al. [100]).

To compare the performance of each algorithm, one often illustrates pseudo regrets of different strategies [97, 47, 98] which are the differences in the true expectations under given strategy and an optimal strategy with perfect information. This is given by

$$\hat{R}(A,T,\Theta) := \sum_{t=1}^{T} \Big(\max_{i \in [K]} h^{(i)}(\Theta) - h^{(A_t)}(\Theta)\Big) \tag{4.3.1}$$

where (A_t) is the observed choices of the given strategy and $h^{(i)}(\Theta) := \mathbb{E}_\Theta[R^{(i)}(Y^{(i)})]$ is the expected reward of the ith option given the true parameter Θ.

In our simulation, we will sample Θ for each simulation but will consider the same Θ's for all algorithms. We illustrate the pseudo regret quantity in terms of its average and quantiles over all simulations.

4.3.1 $1\frac{1}{2}$ bandit problem

We will first illustrate the performance of the ARC algorithm and other algorithms described in Chapter 2 in a simple bandit problem where we have only one unknown parameter, as was considered in Section 4.2. We also include the optimal solution through the discount problem in our simulation and name it 'the Bellman decision.'

We generate the 'true' parameter Θ independently from $N(0, 1)$ and generate 1 trial on the unknown arm as initial information. We run 10^3 Monte-Carlo simulations over $T = 100$ trials. We will use an improper prior as an initial prior for this simulation.

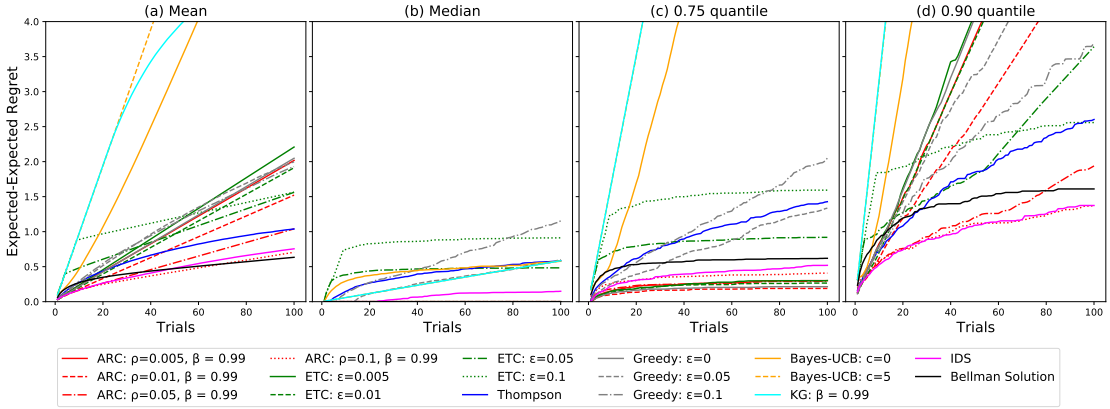

Figure 4.3: Cumulative expected-expected pseudo regret for a $1\frac{1}{2}$ bandit

One can see the Expected-expected regret plot in Figure 4.3 that the ARC algorithm, Thompson sampling, and IDS are compatible with an optimal decision. On the other hand, Bayes-UCB and Knowledge Gradient algorithms often spend too much time exploring and thus suffer from a large regret over this short horizon scenario, whereas greedy and ETC algorithms have adequate performance in this setting.

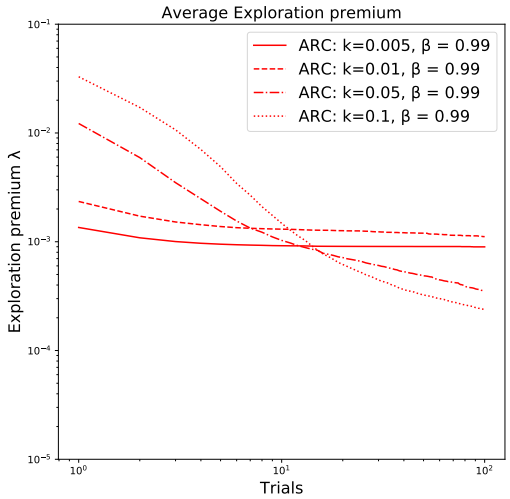

Figure 4.4: Evolution of $\lambda_\rho(m, d)$ in log-log scale

In Section 3.4, we introduced the quantity λ as an additional reward to encourage exploration. We can see in Figure 4.4 that when the value of ρ is large, the value of λ starts higher and collapses faster. This means that when ρ is higher, our algorithm encourages exploration earlier, which makes it require less exploration in the future.

Classical Multi-Armed bandit problem

We now consider a classical bandit model (Section 3.3.1) with 20 arms where $R^{(i)}(y) = \Phi(y)$ for all i. We assume that we only observe the information from the arm we play, i.e. we set the inverse variance $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$. This setting gives our algorithm the least scope to excel, as there is no interaction between arms. To provide a wide range of scenarios in which our strategies must perform, in each simulation, we again generate the 'true' parameter Θ_i independently from $N(0, 1)$. We also generate 5 trials on each arm to provide initial information. (This is required to ensure that Theorem 3.14 holds early in our decision process as discussed in Remark 3.15.) We run 10^3 Monte-Carlo simulations over $T = 5 \times 10^3$ trials. We then consider each algorithm with an improper prior when implementing our algorithm. (i.e., we assume that we have no information about the true prior.)

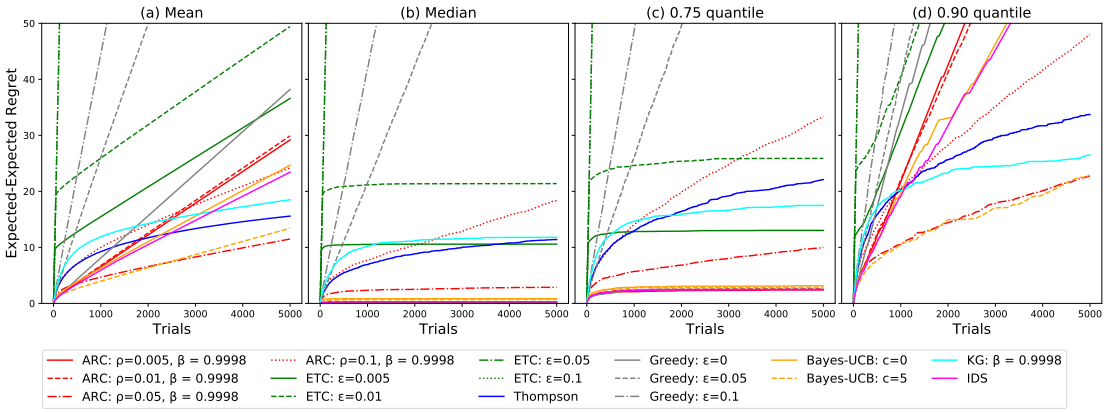

Figure 4.5: Cumulative expected-expected pseudo regret for a classical bandit

Figure 4.5 shows the mean, median, 0.75 quantile and 0.90 quantile of cumulative regret. When the value of ρ is low (e.g., $\rho = 0.005$ or 0.01), our algorithm exploits information early and terminates on one particular arm. This is seen in the median and 0.75 quantile plots, which are flat after a few trials. However, as they terminate early, they are more likely to terminate in a suboptimal decision and suffer from a larger regret on a rare event. This can be seen in the 0.90 quantiles. In contrast, when the value of ρ is high (e.g. $\rho = 0.05$ or 0.1), our algorithm explores longer and lessens the probability of very high regret. This trade-off can be seen in the mean plot, where the regret of our algorithm when $\rho = 0.05$ is lower than all candidates. Let consider the histogram of the terminal regret in Figure 4.6. We can see that when ρ is high, the terminal regret concentrates on small but non-zero values. In contrast, most algorithms, except the Thompson and KG algorithms, have a noticeable dispersion in the tail for the regret. In particular, allowing variation of the value of ρ, we can control the dispersion in the tail distribution of our regret.

Remark 4.5. In Figure 4.5, the average regret of most algorithms (except Thompson and Knowledge Gradients) looks like a linear function. This superficially contrasts with the theoretical guarantees in Kaufmann et al. [64] and in Kirschner and Krause [67]. However, in [64, 67], they consider a frequentist regret (where the optimal order is $\log T$, proved by Lai and Robbins [70]). The constant of the optimal order depends on the true parameter Θ. However, in our simulation, Θ varies between simulations. We cannot expect $\log T$ as an asymptotic regret. For Thompson Sampling, Russo and Van Roy [96] shows that it suffers from a 'Bayesian regret' of order $\sqrt{T}$ when the parameter Θ is allowed to vary, which we observe.

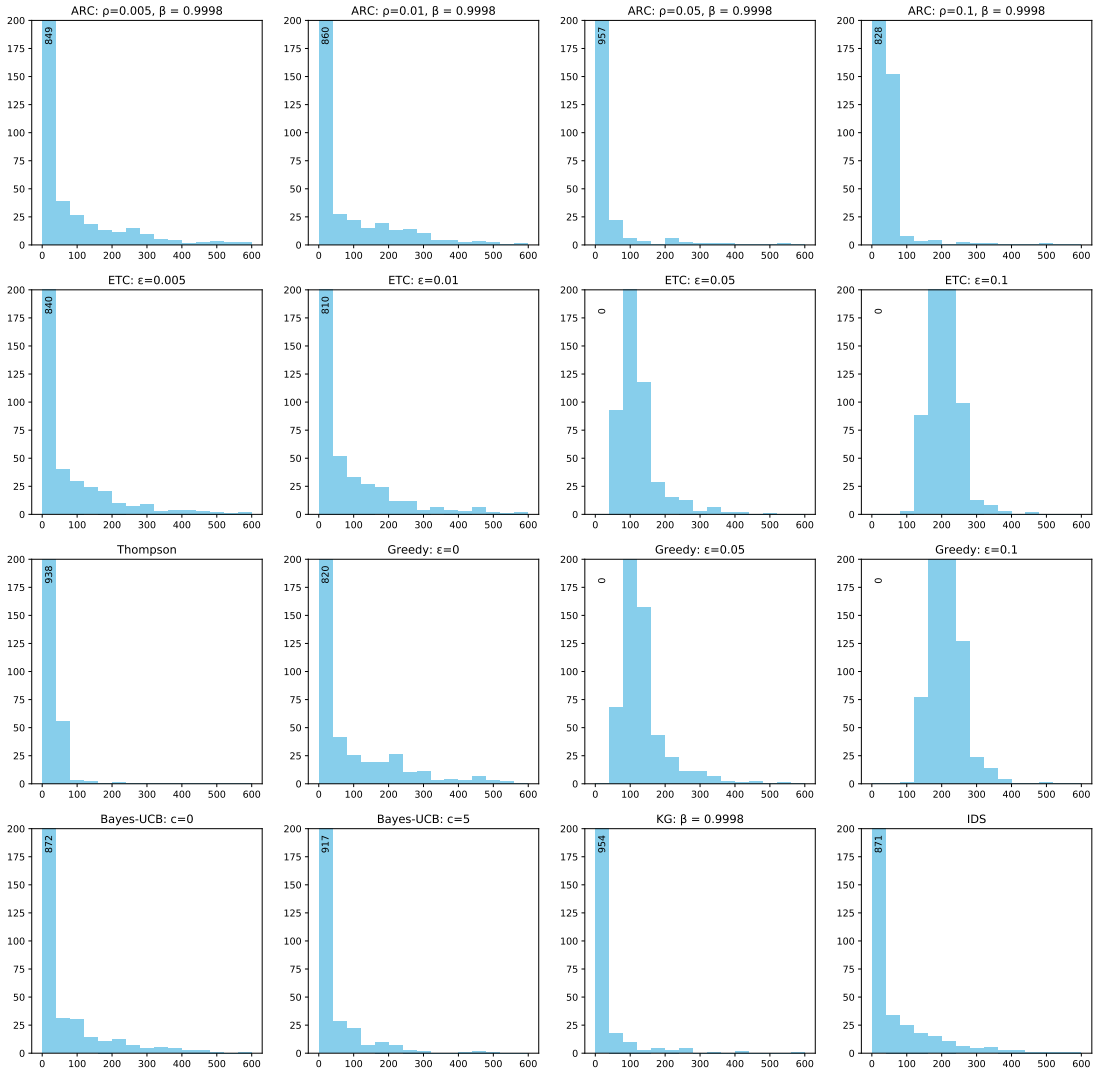

Figure 4.6: Histogram of total regret after 5×10^3 trials (truncated vertically) for a classical bandit

Multi-Armed bandit with a single-arm giving information about others

We consider the case when a single-arm (say $i = 1$) provides substantial information about others, but this arm is more expensive.

In particular, we consider a model in Section 3.3.1 with 20 arms where the reward and the precision matrix p are given by

$$R^{(i)}(y) := \begin{cases} \Phi(y) - 1 & : i = 1 \\ \Phi(y) & : i \neq 1 \end{cases} \qquad \text{and} \qquad p_{ij} = \begin{cases} 5 & : i = 1 \\ 1 & : i = j \text{ but } i \neq 1 \\ 0 & : \text{otherwise.} \end{cases}$$

We again generate the 'true' parameter Θ_i independently from $N(0, 1)$ and generate 5 trials on each arm to provide initial information. We again run 10^3 Monte-Carlo simulations over $T = 5 \times 10^3$ trials and set the discount factor (for our ARC algorithm and the KG algorithm) $\beta = 1 - 1/T$. We then consider each algorithm with an improper prior as our initial information.

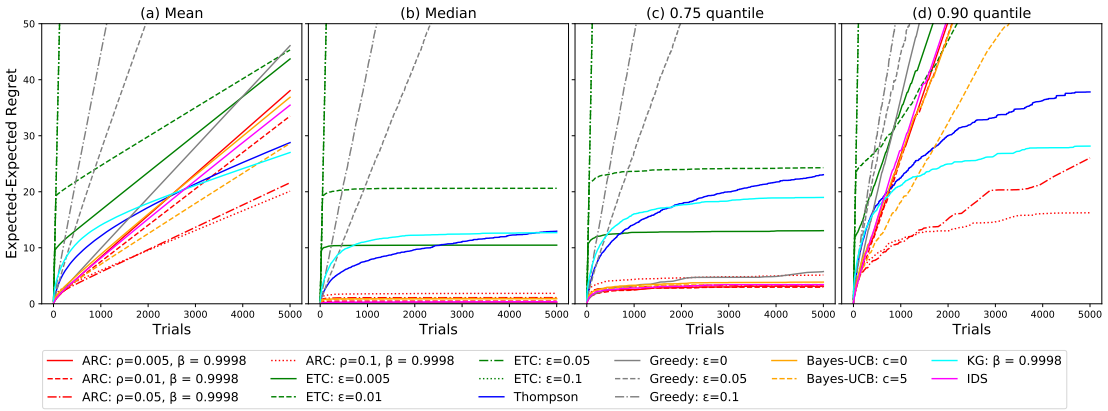

Figure 4.7: Cumulative expected-expected pseudo regret for a bandit with an informative but costly arm

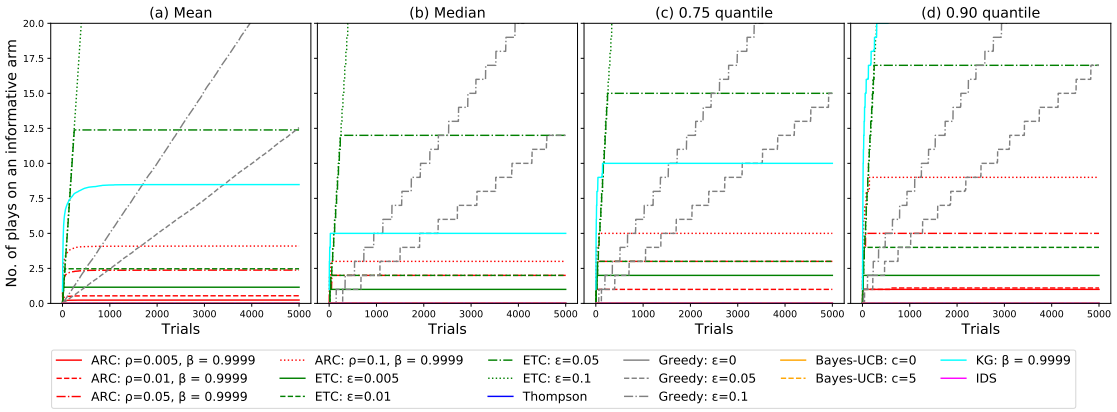

Figure 4.8: The number of times that the informative arm is played

In Figure 4.8, we can see that the greedy (with positive ϵ), the ETC, the ARC, and the KG algorithms are the only algorithms that exploit the expensive but informative arm. We know that the greedy and the ETC algorithm play the first arm due to their random exploration strategy. Nonetheless, plays on the informative arm do not improve these strategies much, as we still see in Figure 4.7, a relatively high regret. In contrast, we know that the ARC and the KG algorithm choose this informative option due to its information. We see in Figure 4.8 that when ρ is high, the ARC algorithm chooses the first option more and exploits this extra information to achieve a low regret. We can see that it outperforms all competitors, including the KG algorithm (which also attempts to exploit the additional information). Our regrets are consistently low up to a high quantile.

Correlated bandit

In this example, we will consider the model when the information provided about other arms is implicit.

We will consider a linear bandit, as in Section 3.3.2. We assume that we have 10 arms, $R^{(i)}(y) = \Phi(y)$ and we only observe the reward from playing, i.e. $l = 1$. Our arms are arranged in a circular network, with each arm being correlated with its neighbours.

We assume that each of our arms depends on the sum of two components of a 10-dimensional vector. In particular, we set $P_i = 1$ for all i and we have $b_i = e^{(i)} + e^{(i+1 \bmod 10)}$. In each simulation, we sample $\Theta \sim N(0, I_{10})$ as a true parameter

and generate 5 trials on each arm to provide initial information. We run 10^3 Monte-Carlo simulation over $T = 5 \times 10^3$ trials. We will consider $\Sigma_0 = I_{10}$ as a (known) prior to implement the algorithms.

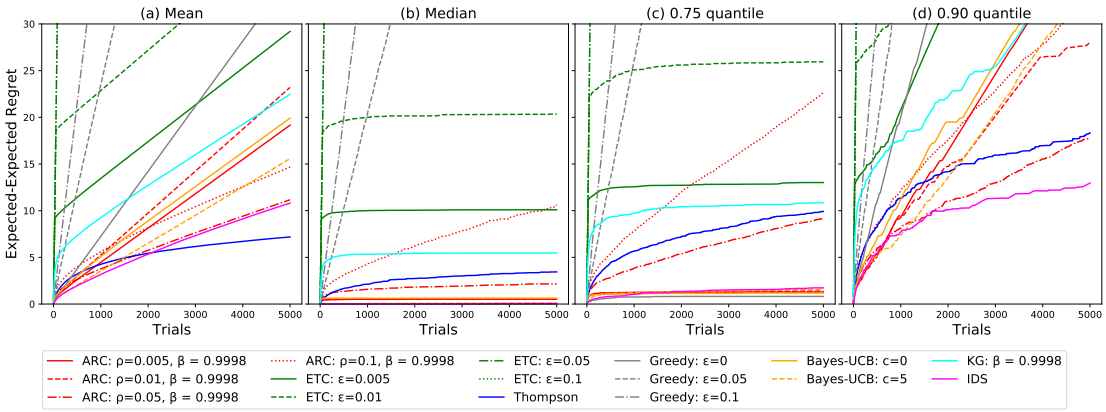

Figure 4.9: Cumulative expected-expected pseudo regret for a correlated bandit

In Figure 4.9, we again see that our algorithms, with a large value of k, give sufficient exploration to the system and hence achieve a low regret on average, as do IDS and Thompson sampling. In contrast, when k is small, our algorithm starts to exploit early and thus obtains a low regret with high probability, but not on average.

Chapter 5

Generalised Linear (batched) Bandit for Dynamic Pricing via the ARC algorithm

The Asymptotic Randomised Control (ARC) algorithm provides a rigorous approximation to the optimal strategy for a broad class of bandit problems while retaining reasonable computational complexity. The algorithm is guaranteed to asymptotically optimise the expected discounted payoff, with error depending on the initial uncertainty of the bandit.

We have already discussed the theoretical results of the ARC algorithm in Chapter 3 and the empirical performance in Chapter 4. Nonetheless, one of the critical limitations of the ARC algorithm is that it requires the distribution of the observation $Y^{(i)} \sim \mu_\Theta^{(i)}$ to admit prior-posterior conjugate in Θ for all i. This requirement is very restrictive for a practical application when we wish to vary the distribution of the observations.

This chapter will discuss how one can apply the ARC algorithm to a more general framework where the observations arrive from a class of Generalised Linear Models (GLMs) in statistics. We will assume that observations came in batches, allowing the large sample theory to approximate the posterior's dynamic via the Kalman filtering equation. We then use this model to run an experiment on dynamic pricing.

The arguments provided in this chapter are not mathematically rigorous, but we focus more on applying the ARC algorithm in practice. This chapter proceeds as follows. In Section 5.1, we consider a generalised linear bandit model and describe how we can use large sample theory and Bayesian statistics to approximately propagate our beliefs. We then outline the implementation of the ARC algorithm in Section 5.2. Finally, in Section 5.3, we use experimental data from [40] to illustrate the

performance of the ARC and other algorithms discussed in Chapter 2 for the Dynamic Pricing problem.

This chapter is based on the paper [36].

5.1 Generalised Linear Bandit

We have already introduced the multi-armed bandit problem as the problem where we face the trade-off between exploration and exploitation. In many literature on bandits, one often looks to establish the regret rate (1.4.2) or (1.4.4). For this reason, the majority of research on bandits focuses in the case when there is no distinction between rewards and observations, i.e. $\sigma\Big(\zeta_1, R^{(A_1)}(Y^{(A_1)}), ..., \zeta_t, R^{(A_t)}(Y^{(A_t)})\Big) = \sigma(\zeta_1, Y_1^{(A_1)}, ..., \zeta_t, Y_t^{(A_t)})$. Therefore, one often makes an assumption directly on the distribution of $R^{(k)}(Y^{(k)})$ rather than on $Y^{(k)}$. This assumption is restrictive and unnatural to model learning in many situation.

For example, in a dynamic pricing problem [40, 79], we want to fix a single product's price from a finite set of prices $\{c_1, ..., c_K\}$ to maximise our revenue. We know that when the price is high, the demand $p(c_k)$ (which we interpret as the chance that each customer will buy the product) is low, but each sale yields a higher return. The agent's reward on a given day is $c_k S_k$ where $S_k|N \sim B(N, p(c_k))$ is the number of customers who buy the product and N is the number of customers on a particular day. On each day, the agent wishes to choose c_k to maximise $c_k \mathbb{E}S_k = c_k p(c_k)\mathbb{E}(N)$. Unfortunately, the agent does not know the true demand $p(c_k)$, and needs to infer it over time.

We expect that there should be some correlation between the true demand $p(c_k)$ for different k. This is because when the price c_k increases, we expect the demand $p(c_k)$ to decrease. How should we model this correlation and fit the pricing problem into a multi-armed bandit framework?

In September 2015, Dibé and Misra [40] ran an experiment, in collaboration with the business-to-business company ZipRecruiter.com, to choose an optimal price in an online sales problem. Their experiment ran in two stages: first collecting data using randomly assigned prices, and then testing their optimal price. A result from their experiment in the first stage is displayed in Figure 5.1.

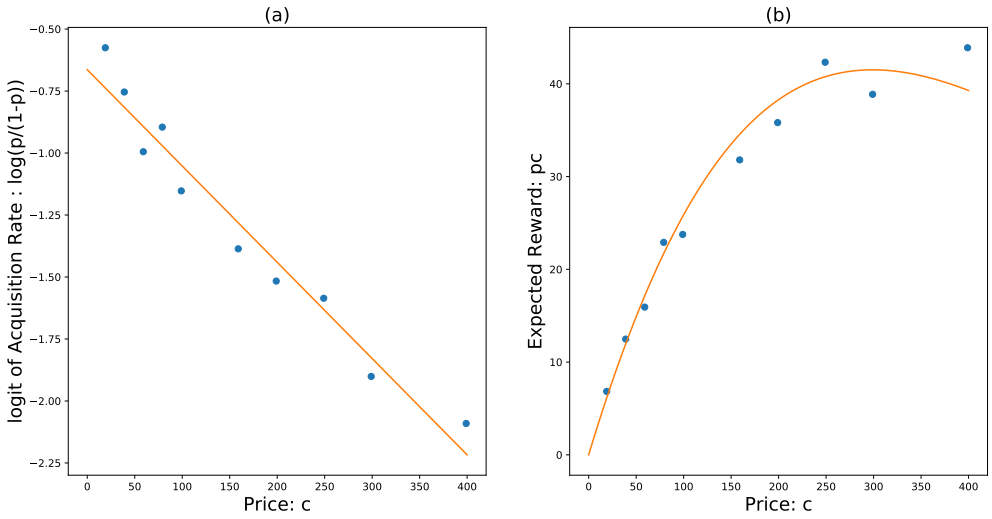

Figure 5.1: (a) Logit of Acquisition Rate, (b) Expected reward per customer.

Figure 5.1(a) displays the relation between the logit (logit := $\log\left(\frac{p}{1-p}\right)$) of the acquisition rate (the proportion of the customers who subscribe), together with its best fit line. The expected reward for each customer ($c_k p(c_k)$) generated by the best fit line and the observed data is illustrated in Figure 5.1(b).

Guided by to Figure 5.1, it is reasonable to consider a logistic model for the probability of subscription; the reward, however, does not fit as naturally into a GLM framework. Suppose we model our demand by the logistic regression. The next question is how to use this model to choose prices sequentially. This requires us to combine the multi-armed bandit problem with generalised linear regression.

5.1.1 Generalised Linear Bandit Model

In this section, we will discuss the learning problem via dynamic pricing. We then extend this framework to a more general setting where observations are sampled from an exponential family whose parameter depends on our decision.

At the beginning of each day, we need to choose a price from the set $\{c_1, ..., c_K\}$. On day t, with chosen price c_k, we observe $N_t^{(k)}$ customers arriving at the store. In order to capture the relations between demands at different prices, we suppose that the probability that each customer buys the product can be modeled by a logistic model, i.e. the relation between the demand $p(c_k)$ and the price c_k is given by

$$\text{logit}\left(p(c_k)\right) = \Gamma_0 + \Gamma c_k = (\Gamma_0, \Gamma)^\top (1, c_k) =: \Theta^\top x^{(k)}$$

where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. The parameter $\Theta = (\Gamma_0, \Gamma)$ is unknown. At the end of day t, we observe $Y_t^{(k)} := \left(N_t^{(k)}, \left\{Q_{i,t}^{(k)}\right\}_{i=1,\dots,N_t^{(k)}}\right)$, where $Q_{i,t}^{(k)}$ takes values in $\{0,1\}$ and indicates whether the product is bought by the ith customer, and collect a reward $R^{(k)}\left(Y_t^{(k)}\right) = c_k \sum_{i=1}^{N_t^{(k)}} Q_{i,t}^{(k)}$.

More generally, we can fit the dynamic pricing problem into the generalised linear model (GLM) framework which is a classical framework in statistics to study parametric model. Let $\{x^{(1)}, ..., x^{(K)}\} \subseteq \mathbb{R}^l$ be a collection of features to be chosen, each corresponding to a choice $\{1, ..., K\}$, and let Θ be a random variable taking values in $\mathbb{R}^l$ representing an unknown parameter. After choosing $x^{(k)}$ at time t, the agent observes $N_t^{(k)}$ independent random variables $\left(Q_{i,t}^{(k)}\right)_{i=1}^{N_t}$ where

$$Q_{i,t}^{(k)} \Big| \Theta \sim_{IID} h(q) \exp\left(\Phi^{(k)} q - G\left(\Phi^{(k)}\right)\right), \tag{5.1.1}$$

$\Phi^{(k)} = \phi\left(\Theta^\top x^{(k)}\right)$ and $Q_{i,t}^{(k)}$ is independent of $N_t^{(k)}$. Here, h, ϕ and G are known functions. For simplicity, we will assume that for each fixed k, the processes $\left(N_t^{(k)}\right)_{t=1}^{\infty}$ are IID with known distribution. After observing $Q_{i,t}^{(k)}$ and $N_t^{(k)}$, we obtain a reward $R^{(k)}(Y_t^{(k)}) = R^{(k)}\left(N_t^{(k)}, Q_{1,t}^{(k)}, ..., Q_{N_t^{(k)},t}^{(k)}\right)$. The objective of our problem is to maximise the total reward in some aspect and thus we can fit this into the multi-armed bandit framework described in Section 1.1.1.

Remark 5.1. Similar frameworks are considered in Filippi et al.[47] and Rusmevichientong and Tsitsiklis [94] but the rewards (instead of the observations) are assumed to be generated from a generalised linear model, which is an unnatural assumption to consider for dynamic pricing or many frameworks in general.

Remark 5.2. A straightforward extension is to allow the observation Q to be such that $T(Q)$ belongs to an exponential family for some function T in an arbitrary dimension. This extension follows the same analysis and will allow us to use the same model to consider, e.g., pricing for multiple products.

5.1.2 Approximate posterior update

To implement a multi-armed bandit algorithm, we need an efficient way to update and record our estimate of the parameter Θ, together with its precision. As our observations are obtained in batches, we shall use a large sample approximation to update via a normal-normal conjugate model.

From (5.1.1), the mean and variance of Q given parameter $\Theta = \theta$ are

$$\mathbb{E}_\theta\left(Q_{i,t}^{(k)}\right) = G'\left(\phi^{(k)}\right) \text{ and } \text{Var}_\theta\left(Q_{i,t}^{(k)}\right) = G''\left(\phi^{(k)}\right).$$

where $\phi^{(k)} = \phi(\theta^\top x^{(k)})$.

Suppose that the link function ϕ is invertible and differentiable. If our model is non-degenerate, G' must also be invertible. We define the *link function* $\psi := (G' \circ \phi)^{-1}$. It then follows from the Central Limit Theorem and the Delta method that

$$\sqrt{n}\big(\Psi_n - \theta^\top x^{(k)}\big) \xrightarrow{d} N\Big(0, 1/[G''\big(\phi^{(k)}\big)\phi'(\theta^\top x^{(k)})^2]\Big) \quad \text{as} \quad n \to \infty, \tag{5.1.2}$$

where $\Psi_n = \psi\Big(\frac{1}{n}\sum_{i=1}^n Q_{i,t}^{(k)}\Big)$. Moreover, by Slutsky's lemma, $\sqrt{nV_n}\big(\Psi_n - \theta^\top x^{(k)}\big) \xrightarrow{d} N(0,1)$, where $V_n := V(\Psi_n) := G''(\phi(\Psi_n))\big(\phi'(\Psi_n)\big)^2$.

When n is not large, $\psi(\frac{1}{n}\sum_{i=1}^n Q_{i,t}^{(k)})$ may not be well-defined for some values of Q. This is the case for the logistic model when $\frac{1}{n}\sum_{i=1}^n Q_{i,t}^{(k)} \in \{0,1\}$. In order to avoid this degeneracy, if M_{t-1} is our running estimate of θ, we consider a linear expansion of ψ around $\psi^{-1}(M_{t-1}^\top x^{(k)})$, which approximates the expected value of $Q_{i,t}^{(k)}$. This approach was used by Fahrmeir [46] to derive an extended Kalman filter with GLM observations, as in (5.1.3).

Suppose that the posterior of Θ at time $t-1$ is

$$\Theta | \mathcal{F}_{t-1}^U \sim N(M_{t-1}, \Sigma_{t-1}).$$

Then, after observing $\big\{Q_{i,t}^{(k)}\big\}_{i=1,\dots,N_t^{(k)}}$, the posterior can be approximately updated by the Kalman filter equations

$$\left.\begin{aligned} M_t &= \Sigma_t\Big(\Sigma_{t-1}^{-1} M_{t-1} + S_t^{(k)} \Psi_t^{(k)} x^{(k)}\Big), \\ \Sigma_t &= \Big(\Sigma_{t-1}^{-1} + S_t^{(k)}\big(x^{(k)}\big)\big(x^{(k)}\big)^\top\Big)^{-1}, \\ \Psi_t^{(k)} &= M_{t-1}^\top x^{(k)} + \big(P_t^{(k)} - \hat{P}_{t-1}^{(k)}\big)\psi'\big(\hat{P}_{t-1}^{(k)}\big), \end{aligned}\right\} \tag{5.1.3}$$

where $S_t^{(k)} := N_t^{(k)} V(\Psi_t^{(k)})$, $\hat{P}_{t-1}^{(k)} := \psi^{-1}(M_{t-1}^\top x^{(k)})$, and $P_t^{(k)} := \sum_{i=1}^{N_t^{(k)}} Q_{i,t}^{(k)} / N_t^{(k)}$.

By Woodbury identity,

$$(A^{-1} + xx^\top)^{-1} = A - \Big(\frac{1}{1 + x^\top A x}\Big) A x x^\top A.$$

Thus, (5.1.3) simplifies to

$$\left.\begin{aligned} M_t &= M_{t-1} + R_t^{(k)}\Big(\Psi_t^{(k)} - M_{t-1}^\top x^{(k)}\Big)\Sigma_{t-1}\left(x^{(k)}\right), \\ \Sigma_t &= \Sigma_{t-1} - R_t^{(k)}\Sigma_{t-1}\left(x^{(k)}\right)\big(x^{(k)}\big)^\top \Sigma_{t-1}, \end{aligned}\right\} \tag{5.1.4}$$

where $R_t^{(k)} = S_t^{(k)} \Big/ \Big(S_t^{(k)} (x^{(k)})^\top \Sigma_{t-1}\left(x^{(k)}\right) + 1\Big)$.

5.2 Implementation of the ARC algorithm

One of the strong assumptions for the ARC algorithm is the posterior-prior conjugate requirement for Θ and the observations $Y^{(k)}$, for all k.

In the generalised linear model, the dependence between the observation $Y^{(k)}$ and the parameter Θ appears through the link function ψ. When ψ is not linear, Θ and $(Y^{(k)})$ may not generally be a posterior-prior conjugate.

We recall that the objective of the ARC algorithm is to maximise

$$\tilde{V}(\beta, U) := \mathbb{E}\Big[\sum_{t=1}^{\infty} \beta^{t-1} f(A_t, M_{t-1}, \Sigma_{t-1})\Big]^1, \tag{5.2.1}$$

where (M_t, Σ_t) evolves with an appropriate dynamic and (A_t) is the corresponding action of the decision (U_t).

This objective is slightly different from the total discounted reward for the bandit problem where we want to maximise

$$V(\beta, U) := \mathbb{E}\Big[\sum_{t=1}^{\infty} \beta^{t-1} R^{(A_t)}(Y_t^{(A_t)})\Big]. \tag{5.2.2}$$

In the case when we have a posterior-prior conjugate pair, with a represented dynamic (M_t, Σ_t), we can write

$$\mathbb{E}\Big[R^{(A_t)}(Y_t^{(A_t)})|\mathcal{F}_{t-1}^U\Big] = f(M_{t-1}, \Sigma_{t-1}, A_t) \tag{5.2.3}$$

for some function f. This yields the equality between (5.2.1) and (5.2.2).

In the GLM framework, the representation (5.2.3) does not necessary holds. However, when the batch size is large, it follows from large sample theories that $\text{Law}(\Theta|\mathcal{F}_{t-1}^U) \approx N(M_{t-1}, \Sigma_{t-1})$. In particular,

$$\mathbb{E}\Big[R^{(k)}(Y_t^{(k)})|\mathcal{F}_{t-1}^U\Big] \approx \mathbb{E}\Big[R^{(k)}(Y_t^{(k)})|\Theta \sim N(M_{t-1}, \Sigma_{t-1})\Big]. \tag{5.2.4}$$

Hence, we may consider the RHS of the expression above as our function $f_k(m, \Sigma)$ and assume that the posterior dynamic follow (5.1.4). We can now apply the ARC algorithm to $\tilde{V}$ instead of V.

For the convenience of the reader, we will recall the analogy of the ARC algorithm. The key idea of the ARC algorithm is to estimate the optimal solution to (5.2.1) via a Markov decision process with (m, Σ) as an underlying state. A smooth approximation is obtained by introducing a preference for random decisions in the objective

[1] we simply consider the variance Σ_t as our precision part (instead of D_t as in Chapter 3, see Remark 3.7 for the justification).

function (5.2.1), in particular adding a reward $\lambda\mathcal{H}(A_t)$ to $f^{(A_t)}$ in (5.2.1), where $\mathcal{H}$ is a smooth entropy function (e.g. Shannon entropy, which we use here). The scale of this preference is controlled through the parameter λ, which is determined dynamically in order to have a negligible effect when uncertainty is low. This approximation results in a semi-index strategy, which amounts to computing the solution $a \in \mathbb{R}^K$ to the fixed point equation:

$$a = f + \Big(\frac{\beta}{1-\beta}\Big)L^{\lambda}(a)$$

where the term f corresponding to expected rewards over one time step (quantifying the gain from immediate exploitation), and $L^{\lambda}(a)$ is an exploration term. Here, we will use (5.2.4) as an approximation to the expected rewards.

The solution a shall be interpreted as measuring the immediate reward and the increase in total reward arising from each choice, taking into account the effect of learning on future rewards. The entropy term results in the ARC algorithm applying a softmax function to a, yielding conditional probabilities of choosing each arm rather than a deterministic choice.

The ARC algorithm's implementation requires the computation of the dynamics of the posterior parameter and the derivative of the expected one-period reward with respect to the underlying state (i.e., the posterior parameters). In the GLM framework, we have introduced (5.1.4) to simplify our posterior dynamic, allowing us to estimate the relevant terms in the ARC algorithm for this case. A possible implementation can be given by the following procedures:

Step I: Estimate the (expected) dynamics of the posterior parameter. The ARC algorithm requires the computation of how we expect the parameter estimate and its precision to change. In the case of interest, we will treat the posterior mean m and variance Σ of the parameter Θ as the 'estimator' and 'precision' in the ARC algorithm.

Let m and Σ be the posterior mean and variance conditional on $\mathcal{F}_t^U$ and let $M^{(k)}$ and $\Sigma^{(k)}$ be their update after we choose the kth arm. As discussed in Section 5.1.2, we can update $M^{(k)}$ and $\Sigma^{(k)}$ by (5.1.4);

$$M_t = M_{t-1} + R_t^{(k)}\Big(\Psi_t^{(k)} - M_{t-1}^{\top}x^{(k)}\Big)\Sigma_{t-1}\left(x^{(k)}\right),$$
$$\Sigma_t = \Sigma_{t-1} - R_t^{(k)}\Sigma_{t-1}\left(x^{(k)}\right)\left(x^{(k)}\right)^{\top}\Sigma_{t-1},$$

where $\Psi_t^{(k)} = M_{t-1}^{\top}x^{(k)} + \big(P_t^{(k)} - \hat{P}_{t-1}^{(k)}\big)\psi'\big(\hat{P}_{t-1}^{(k)}\big)$, $R_t^{(k)} = S_t^{(k)}\Big/\Big(S_t^{(k)}(x^{(k)})^{\top}\Sigma_{t-1}\left(x^{(k)}\right)+1\Big)$, $S_t^{(k)} := N_t^{(k)}V(\Psi_t^{(k)})$, $\hat{P}_{t-1}^{(k)} := \psi^{-1}(M_{t-1}^{\top}x^{(k)})$, and $P_t^{(k)} := \sum_{i=1}^{N_t^{(k)}} Q_{i,t}^{(k)}/N_t^{(k)}$.

From (5.1.2), $\sqrt{n}\big(\Psi_n - \theta^\top x^{(k)}\big) \xrightarrow{d} N\Big(0, 1/[G''\big(\phi^{(k)}\big)\phi'(\theta^\top x^{(k)})^2]\Big)$. By considering the Slutsky's lemma, we can estimate $\Psi^{(k)}|\Theta \approx N\big(\Theta^\top x^{(k)}, 1/S^{(k)}\big)$. Assuming the posterior $\Theta \sim N(m,\Sigma)$ gives $\Theta^\top x^{(k)} \sim N\Big(m^\top x^{(k)}, \big(x^{(k)}\big)^\top \Sigma \, \big(x^{(k)}\big)\Big)$. Hence, the conditional distribution of $\Psi^{(k)}|(m,\Sigma)$ can be approximated by $\mathrm{Law}(\Psi^{(k)}|m,\Sigma) \approx N\Big(m^\top x^{(k)}, \big(\frac{S^{(k)}(x^{(k)})^\top \Sigma \, (x^{(k)})+1}{S^{(k)}}\big)\Big)$.

Therefore, (5.1.4) yields an approximate innovation representation

$$\Delta M^{(k)} \approx \left(\frac{S^{(k)}}{S^{(k)}\big(x^{(k)}\big)^\top \Sigma \, \big(x^{(k)}\big) + 1}\right)^{1/2} \Sigma \, \big(x^{(k)}\big) Z,$$

where $Z \sim N(0, I)$.

As $\Theta \sim N(m,\Sigma)$, when Σ is small, we may estimate $S^{(k)} \approx n_k V\big(m^\top x^{(k)}\big)$ where $n_k = \mathbb{E}\big(N^{(k)}\big)$.

Hence, the dynamics of our state (m,Σ) can be approximated by

$$\begin{aligned}
\mathbb{E}_{m,\Sigma}\big(\Delta M^{(k)}\big) &\approx \tilde{\mu}^{(k)}(m,\Sigma) := 0,\\
\mathrm{Var}_{m,\Sigma}\big(\Delta M^{(k)}\big) &\approx \big(\tilde{\sigma}\tilde{\sigma}^\top\big)^{(k)}(m,d) := w(m,\Sigma,k)\Sigma \, \big(x^{(k)}\big) \, \big(x^{(k)}\big)^\top \Sigma,\\
\mathbb{E}_{m,\Sigma}\big(\Delta \Sigma^{(k)}\big) &\approx \tilde{b}^{(k)}(m,d) := -w(m,\Sigma,k)\Sigma \, \big(x^{(k)}\big) \, \big(x^{(k)}\big)^\top \Sigma,
\end{aligned}$$

where $w(m,\Sigma,k) := \big(\frac{n_k V(m^\top x^{(k)})}{n_k V(m^\top x^{(k)})(x^{(k)})^\top \Sigma \, (x^{(k)})+1}\big)$.

Step II: Compute the expected reward f and learning function L^λ using the estimated dynamics. We next compute the expected reward given the (estimate) posterior parameter, that is $f : \mathbb{R}^l \times \mathcal{S}_+^l \to \mathbb{R}^K$ with components

$$f_k(m,\Sigma) := \mathbb{E}\Big(R^{(k)}\big(Y^{(k)}\big)\Big|\Theta \sim N(m,\Sigma)\Big),$$

where $\mathcal{S}_+^l$ is the family of positive definite $\mathbb{R}^{l\times l}$ matrices, and the learning function $L^\lambda : \mathbb{R}^K \times \mathbb{R}^l \times \mathcal{S}_+^l \to \mathbb{R}^K$ with components

$$\begin{aligned}
L_k^\lambda(a,m,\Sigma) := \langle \mathcal{B}^\lambda(a,m,\Sigma); \mathbb{E}_{m,\Sigma}\big(\Delta\Sigma^{(k)}\big)\rangle + \langle \mathcal{M}^\lambda(a,m,\Sigma); \mathbb{E}_{m,\Sigma}\big(\Delta M^{(k)}\big)\rangle\\
+ \frac{1}{2}\langle \Xi^\lambda(a,m,\Sigma); \mathrm{Var}_{m,\Sigma}\big(\Delta M^{(k)}\big)\rangle,
\end{aligned} \tag{5.2.5}$$

where we define

$$\begin{aligned}
\mathcal{B}^\lambda &:= \sum_k \nu_k^\lambda(a)(\partial_\Sigma f_k), \quad \mathcal{M}^\lambda := \sum_k \nu_k^\lambda(a)(\partial_m f_k),\\
\Xi^\lambda &:= \sum_k \nu_k^\lambda(a)(\partial_m^2 f_k) + \frac{1}{\lambda}\sum_{j,k} \eta_{jk}^\lambda(a)(\partial_m f_j)(\partial_m f_k)^\top,\\
\nu_k^\lambda(a) &:= \exp\big(a_k/\lambda\big)/\sum_j \exp\big(a_j/\lambda\big), \qquad \eta_{jk}^\lambda(a) := \nu_j^\lambda(a)\big(\mathbb{I}(j=k) - \nu_k^\lambda(a)\big).
\end{aligned}$$

Since the ARC algorithm focuses on the regime where Σ is small, we may estimate our expected reward by $\tilde{f} : \mathbb{R}^l \to \mathbb{R}^K$ with components

$$f_k(m, \Sigma) \approx \tilde{f}_k(m) := \mathbb{E}_m\left(R^{(k)}\left(N^{(k)}, \left(Q_i^{(k)}\right)_{i=1}^{N^{(k)}} \right) \middle| \Theta = m \right) = h_k\left(m^\top x^{(k)}\right), \qquad (5.2.6)$$

for some function h_k. The last equality follows from the fact that $\{Q_i^{(k)} | \Theta = m\}_{i=1,\dots,\infty} \sim_{IID} g(q; m^\top x^{(k)})$ and $N^{(k)}$ is independent of $\{Q_i^{(k)}\}_{i=1,\dots,\infty}$.

Using the estimates $\tilde{\mu}^{(k)}$, $\left(\tilde{\sigma}\tilde{\sigma}^\top\right)^{(k)}$, $\tilde{b}^{(k)}$, and $\tilde{f}_k$, we can approximate L^λ in the ARC algorithm by

$$\begin{aligned}\tilde{L}_k^\lambda := \frac{1}{2} v(m, \Sigma, k) \Bigg[& \sum_{j=1}^K \nu_j^\lambda(a) h_j''\left(m^\top x^{(j)}\right)\left(r_{kj}(\Sigma)\right)^2 \\ & + \frac{1}{\lambda}\left(\sum_{j=1}^K \nu_j^\lambda(a)\left(h_j'\left(m^\top x^{(j)}\right) r_{kj}(\Sigma)\right)^2 - \left(\sum_{j=1}^K \nu_j^\lambda(a) h_j'\left(m^\top x^{(j)}\right) r_{kj}(\Sigma)\right)^2 \right) \Bigg], \quad (5.2.7)\end{aligned}$$

where $r_{kj}(\Sigma) := \left(x^{(j)}\right)^\top \Sigma \left(x^{(k)}\right)$.

Step III: Use the ARC algorithm for the generalised linear bandit. Finally, to apply the ARC algorithm, one also needs to choose a rescaling function to encourage randomisation when estimates are uncertain. Here, we choose $\lambda_\rho(m, \Sigma) := \rho\|\Sigma\|$, using the Euclidean norm $\|\Sigma\| = \sqrt{\langle \Sigma; \Sigma \rangle}$. We refer to Chapter 3 for further discussion on the effect of this choice and alternatives.

To run the ARC algorithm, we need to choose the parameters $\rho > 0$ and $\beta \in (0, 1)$. Here ρ determines the randomness of our early choices, encouraging early exploration, while β is a discount factor, which is used to value the future reward. The user can choose these parameters. We again choose $\beta = 1 - 1/T$. The reader may also refer to algorithm 1 and 2 in Section 4.1 to recall pseudo-codes for the ARC algorithm and bandit simulation.

5.3 Simulation for Dynamic Pricing

This section will consider a hierarchical model to construct a bandit environment from the observed data. We will then use the data from Dubé and Misra's experiment [40] to construct a bandit environment and test algorithms.

5.3.1 Simulate bandit environment

As the strategy followed determines the observations available, we cannot directly use historical data to test the algorithm. However, we can use the data to construct an environment to run tests. We take a Bayesian view and build a simple hierarchical model (with an improper uniform prior). Effectively, this assumes that our observations come from an exchangeable copy of the world we would deploy our bandits in,

with the same (unknown) realised value of Θ. Then we will use Laplace approximation, as in Russo et al. [98, Chapter 5] or Chapelle and Li [26], to obtain a posterior sample to simulate the markets.

Remark 5.3. It is worth emphasising that when implementing the algorithm in the simulation, we do not assume that our algorithms know the distribution that we use to simulate the parameter, Θ. Instead, the simulations are initialised with an almost uninformative prior.

To construct a posterior for Θ given historical data, we assume that we have a collection of observations $(q_i^{(1)})_{i=1}^{r_1}$, $(q_i^{(2)})_{i=1}^{r_2}$, ..., $(q_i^{(K)})_{i=1}^{r_K}$ from an exponential family modeled by (5.1.1). We denote the corresponding log-likelihood function of Θ by $\ell\big(\theta; \big(q_i^{(k)}\big)\big)$.

Let $\hat{\theta}$ be the maximum likelihood estimator. i.e. $\hat{\theta} = \arg\max_\theta \ell\big(\theta; \big(q_i^{(k)}\big)\big)$. Then we may approximate the log-likelihood function by

$$\ell\Big(\theta; \big(q_i^{(k)}\big)\Big) \approx \frac{1}{2}(\theta - \hat{\theta})^\top \partial_\theta^2 \ell\big(\hat{\theta}; \big(q_i^{(k)}\big)\big)(\theta - \hat{\theta}) + c,$$

for c which does not depend on θ. Therefore, starting from the uninformative (improper, uniform) prior, we can estimate the posterior of the parameter Θ by

$$\Theta \sim N\Big(\hat{\theta}, \big(- \partial_\theta^2 \ell\big(\hat{\theta}; (q_i^{(k)})\big)\big)^{-1}\Big). \tag{5.3.1}$$

In particular, if the parameter for the exponential family is given in its canonical form, i.e., when ϕ in (5.1.1) is the identity, then the observed Fisher information is

$$- \partial_\theta^2 \ell\big(\hat{\theta}; (q_i^{(k)})\big) = \sum_{k=1}^{K} r_k G''\big(\hat{\theta}^\top x^{(k)}\big)\big(x^{(k)}\big)\big(x^{(k)}\big)^\top.$$

We will use the simulated values of Θ as the (hidden) realisations to test our algorithms.

5.3.2 Dynamic Pricing Simulation

Finally, we can implement the multi-armed bandit algorithm to run a simulation for the dynamic online pricing problem.

Recall that in this model, at each time t, we need to choose $x^{(k)} = (1, c_k) \in \mathbb{R}^2$ where c_k is the price of the product. We then observe the number of customers $N_t^{(k)}$ and whether each customer buys the product or not in terms of a binary random variable $\big(Q_{i,t}^{(k)}\big)_{i=1}^{N_t^{(k)}}$. The reward that we receive in each step is given by $R^{(k)}\big(N_t^{(k)}, Q_{1,t}^{(k)}, ..., Q_{N_t,t}^{(k)}\big) := c_k \sum_{i=1}^{N_t^{(k)}} Q_{i,t}^{(k)}$. We will model $\big(Q_{i,t}^{(k)}\big)$ by a logistic model, i.e. $G(z) = \log(1 + e^z)$ and $\phi(z) = z$ for the functions ϕ and G in (5.1.1).

We can now write down the expected reward,

$$\mathbb{E}_\theta\Big(R^{(k)}\Big(N_t^{(k)}, Q_{1,t}^{(k)}, ..., Q_{N_t^{(k)},t}^{(k)}\Big)\Big) = n_k c_k G'\big(\theta^\top x^{(k)}\big),$$

where $n_k = \mathbb{E}(N_t^{(k)})$. In particular, the function h_k in (5.2.6) and (5.2.7) is given by $h_k(y) = n_k c_k G'\big(y\big)$.

Market data and simulation environment

In Dubé and Misra [40], in stage one of their experiment, they randomly assigned one of ten experimental pricing cells to 7,867 different customers who reached Ziprecruiter's paywall. The exact numbers of customers assigned to each price are not reported. Hence, we will assume that there are exactly 787 customers for each price. We then use their reported subscription rate to estimate the exact numbers of customers who decided to subscribe given each price.

Using this data, we apply (5.3.1) to infer an approximate posterior distribution:

$$\Theta \sim N\left(\begin{bmatrix} -6.42 \times 10^{-1} \\ -4.03 \times 10^{-3} \end{bmatrix}, \begin{bmatrix} 1.90 \times 10^{-3} & -8.86 \times 10^{-6} \\ -8.86 \times 10^{-6} & 6.82 \times 10^{-8} \end{bmatrix}\right). \tag{5.3.2}$$

To compare performance, we will consider each algorithm over a period of one year (365 days) and only allow the agent to change the price at the end of each day. We assume that a common price must be shown to all customers each day. We will also assume that the chosen price does not affect the number of customers reaching the paywall. i.e. we assume that $N_t^{(k)} \equiv N_t$.

We run 5×10^3 independent simulations of the different market situations, where for each simulation, the parameter Θ is sampled from (5.3.2). We also independently sample $(N_t)_{t=1}^{365} \sim \text{Poisson}(270)$ to represent the number of visitors on each day[2]. The simulated subscription probability and the simulated expected revenue per customer for each price level are illustrated in Figure 5.2 with two standard deviation bands.

Remark 5.4. Misra et al. [79] also used the same data to illustrate dynamic online pricing as a classical multi-armed bandit problem. They tackled this problem using a modification of the classical UCB algorithm and UCB-tuned algorithm [10] where the demand correlation is not taken into consideration.

In our simulation, we will also consider the classical UCB [1] and the UCB-tuned algorithm [10] as candidates. These algorithms are similar to the Bayes-UCB algorithm but consider the problem from a frequentist perspective and ignore the correlation between outcomes. In particular, we do not use (5.1.4) to propagate our belief, but we record the reward of each arm separately. The reader may refer to Chapter 2, Misra et al. [79], Burtini et al. [22] or Auer et al. [10] for the discussion on UCB and UCB-tuned algorithm.

Simulation Results

We apply each algorithm described in Chapter 2 and the ARC algorithm, with $m_0 = (0, 0)$ and $\Sigma_0 = I_2$ as an initial prior for Θ and use (5.1.4) to propagate the prior.

To assess the performance of each algorithm, one often compares the cumulative pseudo-regret of each algorithm given the true parameter Θ:

$$R\big(\theta, T, (A_t)\big) := \sum_{t=1}^{T} \Big(\max_k \tilde{f}_k(\theta) - \tilde{f}_{A_t}(\theta) \Big),$$

[2] In Dubé and Misra, we can observe that ZipRecruiter.com had roughly $8,000$ visitors per month. Hence, it is reasonable to assume roughly 270 visitors per day.

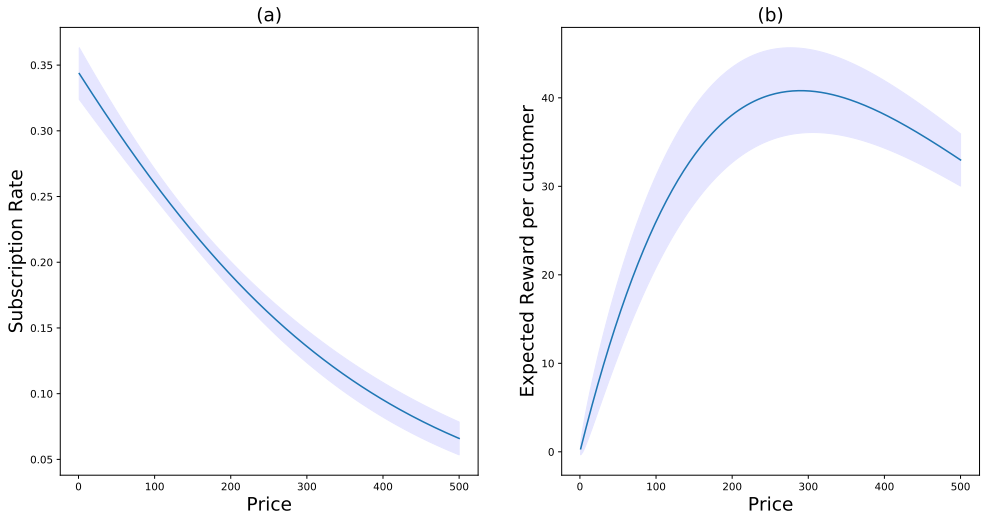

Figure 5.2: (a) Subscription rate, (b) Expected reward per customer.

where $\tilde{f}_k(\theta) := \mathbb{E}\left[R^{(k)}\left(Y_t^{(k)}\right)\middle|\Theta = \theta\right] = nc_k G'\left(\theta^\top x^{(k)}\right)$, $G(y) = \log(1 + e^y)$, (A_t) is the sequence of actions that the algorithm chooses and $n = \mathbb{E}(N_t^{(k)})$.

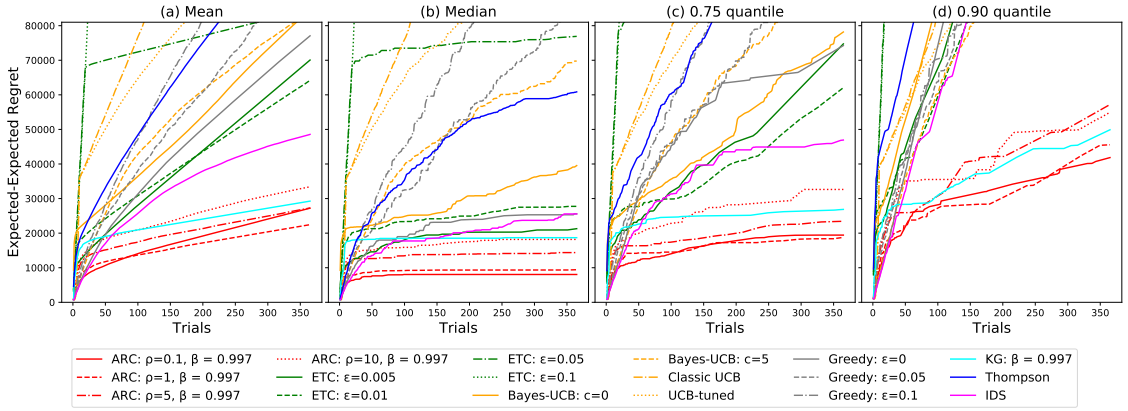

Figure 5.3: Cumulative expected-expected pseudo regret for dynamic pricing

Figure 5.3 shows the mean, median, 0.75 quantile and 0.90 quantile of cumulative pseudo-regret of each algorithm described in Chapter 2. We see that most algorithms outperform the classical UCB and the UCB-tuned used in Misra et al. [79], which is unsurprising as these approaches ignore the correlation between demand at different prices. We also see that the ARC algorithm outperforms all other algorithms both on average and in extreme cases (as shown by the quantile plots).

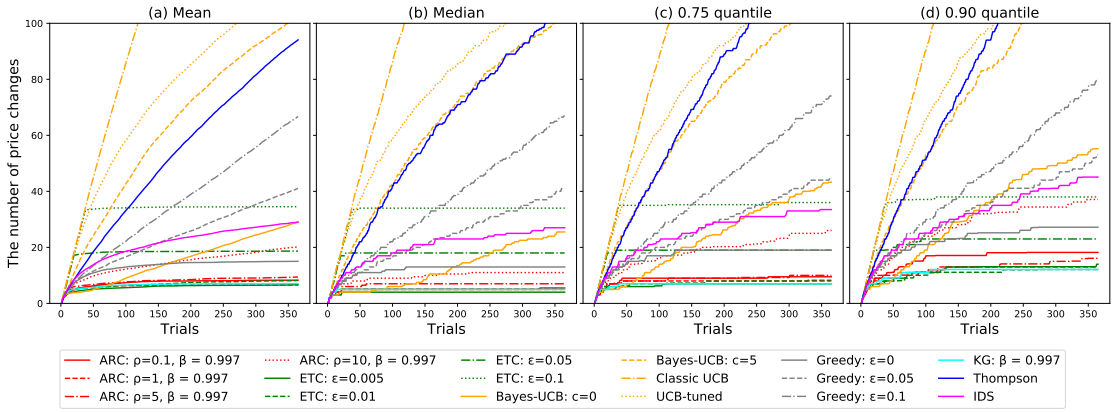

Figure 5.4: The number of price changes

In addition to the regret criteria, when displaying the price, the agent may not want to change the price too frequently. This is why the Explore-Then-Commit (ETC) algorithm is often preferred when considering a pricing problem. Figure 5.4 shows that the ARC algorithm and KG algorithm typically require a small number of price changes but still achieve a reasonably low regret (as shown in Figure 5.3). More commonly used algorithms, e.g., Thompson sampling and the UCB type algorithms require a larger number of price changes.

Chapter 6

Nonlinear expectation, Time-Consistency and Optimality

In earlier chapters, we discussed the ARC algorithm and its application. An intuitive explanation of the ARC algorithm is to quantify the incremental reward in terms of the immediate reward and learning premium. In a sense, we can see the learning premium as being optimistic toward uncertainty. However, as discussed in the introduction of Chapter 3, we, in fact, are not optimistic toward uncertainty. We only act optimistically to uncertainty due to our perception of the value of information and it turns out that these can be regarded as the same in the classical bandit case.

We also discussed in Section 1.5.2 that people often have a strict preference for choices they understand better. In particular, we are generally pessimistic about uncertainty. The aim of this chapter, Chapter 7 and Chapter 8 is to allow bandits to interact with uncertainty aversion, via proving Gittins' theorem under uncertainty aversion. Therefore, we will only focus on the case where we only observe information and collect the pay-off from the arm we play. Here, the observation of each arm does not depend on each other. We shall see this model as the arms are being 'independent'. It is also worth to point out that our proved theorem will not require any Markovian structure in the model. However, the Markovian structure is still useful in a computational aspect.

To capture uncertainty aversion, we consider an objective function based on a 'nonlinear expectation' instead of a classical linear expectation. [1] This is a standard method to handle uncertainty aversion in modern stochastic control. (see e.g. Cohen and Allan [4], Ekren et al. [41], Riedel [90])

We can use the nonlinear expectation to model uncertainty from both Bayesian and frequentist perspectives. Bayesians often chooses their favourite prior and formulate a problem into a probabilistic framework. The 'ambiguity' of the parameter is thus seen in terms of 'risk' (rather than 'uncertainty' discussed in Section 1.5).

[1] We have already considered a smooth 'nonlinear expectation' defined on a finite space in Section 3.4.2 as a tool to smooth our value function. This chapter will explore 'nonlinear expectations' in greater detail and will be used as a tool to model uncertainty aversion that gives the opposite effect from what has been done in Chapter 3.

[2] Nonetheless, under this framework, we can also model uncertainty through our choice of priors. (e.g. [90, 74]). On the other hands, frequentists may treat uncertainty through the confidence interval of the parameter estimate. We can construct our nonlinear expectation to represent uncertainty (e.g. [29]). Our approach proves Gittins theorem under a general nonlinear expectation framework and thus could model uncertainty from both frequentists and Bayesian.

There is another aspect that requires attention when making decision through time under nonlinear expectation. As Keynes is said to have remarked, "When the facts change, I change my mind. What do you do, sir?". The question a rational decision-maker faces is, "if I suspect that I will change my opinions or preferences tomorrow, how do I account for this today?" This is a fundamental aspect of stochastic control theory, which is often seen through the dynamic programming principle and results in the tractability of backward induction. To maintain this idea and mathematical tractability in our control problem, we usually require a time-consistency property in the nonlinear expectation. This property is equivalent to the tower property in the classical expectation.

We have seen in Section 1.1.1 that for a bandit problem, the observations are determined by the strategy we use. This means that to apply the nonlinear expectation framework to the multi-armed bandit problem, we require the nonlinear expectation to be consistent over multiple filtrations. Unfortunately, we cannot guarantee that this will be the case (see Cohen [28] or Graf [57]). Existing theories for a time-consistent nonlinear expectation always assume that the filtration is fixed and thus cannot be directly extended to this framework.

To overcome the prescribed problems, we construct our space such that the nonlinear expectation is consistent in the marginal projection, but not necessarily consistent on the entire space on which the multiple arms are defined. As the constructed nonlinear expectation is not necessarily consistent, we consider a relaxation of the dynamic programming principle and view decisions for the multi-armed bandit from a different perspective to what has been discussed in Section 1.1.1, which allows a natural marginal projection. This way of defining the strategy was modified from an approach proposed by Mandelbaum [77], and only makes sense when the arms are 'independent'.[3] We will leave the detail discussion of this strategic structure until Chapter 7 and will only focus on the time-consistency and optimality in this chapter.

This chapter proceeds as follows. Section 6.1 focuses on defining the probability space for the multi-armed bandit and states the required assumptions. We then discuss the theory of (consistent) nonlinear expectation in Section 6.2 and make a natural extension of the nonlinear expectation to define an operator over multiple filtrations in Section 6.3. Finally, in Section 6.4, we discuss a previous attempt to consider the robust Gittins' theorem. However, they show that the robust index is sub-optimal in a standard robust sense. Thus, we propose an alternative optimal-

[2]The term risk is commonly refer to the situation where we do not know the outcome but know the probability whereas the term uncertainty is often refer to the case when the probability is not known.

[3]In particular, this restricts our consideration to the classical bandit problem or a contextual bandit problem where Gittins' theorem holds, under a classical linear expectation.

ity criteria that connects many used approaches to overcome time-inconsistency in decision-making.

As the nonlinear expectation is often assumed to be convex and is commonly used to consider the cost problem. Hence, for the rest of this thesis, we will consider the minimisation instead of maximisation for notational convenience.

This chapter is based on the paper [35].

6.1 Problem setup and Assumptions

We will first introduce a new framework, which slightly modifies the Mandelbaum [77] and El Karoui and Karatzas [63] settings.

Suppose that we have K arms. The kth arm is associated with a filtered probability space $\big(\Omega^{(k)}, \mathbb{P}^{(k)}, (\mathcal{F}_t^{(k)})_{t\geq 0}\big)$. Playing this arm for the tth time realises a non-negative bounded cost $h^{(k)}(t)$, where the process $\big(h^{(k)}(t)\big)_{t\geq 1}$ is adapted to $(\mathcal{F}_t^{(k)})_{t\geq 0}$. We assume that $\mathcal{F}_0^{(k)} = \{\phi, \Omega^{(k)}\}$.

Remark 6.1. In section 1.1.1, we introduce $Y_t^{(k)}$ as an observation when the kth arm is played at time t. We then represent the filtration of the problem corresponding to the control. However, in the nonlinear expectation setup, we need to avoid the direct dependence between control and filtrations. Thus, we assume that the observations of each arm correspond to an individual filtration. In particular, $\mathcal{F}_t^{(k)}$ represents available information from the kth arm when **the kth arm** was played t time. Roughly speaking, one can see $\mathcal{F}_t^{(k)}$ as $\sigma\big(Y_{s_1}^{(k)}, ..., Y_{s_t}^{(k)} : s_i = A_u \text{ for some } u\big)$.

The cost process $\big(h^{(k)}(t)\big)_{t\geq 1}$ shall be interpreted as $\big(-R^{(k)}(Y_t^{(k)})\big)_{t\geq 1}$ where we simply omit the dependence of $Y_t^{(k)}$ for notational simplicity.

To avoid technical difficulties, we will make two additional technical assumptions on the filtration and the cost processes.

Assumption 6.1. *There exists $T^{(k)} < \infty$ such that $\mathcal{F}_t^{(k)} = \mathcal{F}_{T^{(k)}}^{(k)}$ for all $t \geq T^{(k)}$.*

Assumption 6.1 is required to obtain a simple robust representation (Theorem 6.3) of nonlinear expectation which will be given shortly. This is our main ingredient to define uncertainty in the orthant space of the multiple arms. One may remove this assumption and simply consider a robust representation directly, with the requirement that some pasting property is satisfied.

Assumption 6.2. *For each $k \in [K]$, the process $h^{(k)}(t)$ is uniformly bounded and converges to its bound.*

Assumption 6.2 is purely technical. We may replace boundedness of $h^{(k)}$ by an integrability assumption on the total discounted cost (as in [63]); we then need to generalise the domain of the nonlinear expectation. We can also remove the assumption on the convergence of $h^{(k)}$ to its bound, but a more careful analysis is required to assure that all considered stopping times are a.s. finite. Given the discount factor, the cost decays exponentially fast. This assumption does not have large impact on our modelling but it will be proved to simplify our analysis.

6.2 Nonlinear expectations and Time-Consistency

We have assumed that $\left(\Omega^{(k)}, \mathbb{P}^{(k)}, (\mathcal{F}_t^{(k)})_{t\geq 0}\right)$ represents the information from the kth arm. In this section, we will quantify the (Knightian) uncertainty of each arm separately through 'nonlinear expectation' operators. We will also discuss how we can use these tools to study a control problem under uncertainty while retaining some form of time consistency.

We will use a 'nonlinear expectation' $\mathcal{E}^{(k)}$ to model uncertainty on the space $\left(\Omega^{(k)}, \mathbb{P}^{(k)}, (\mathcal{F}_t^{(k)})_{t\geq 0}\right)$ of a single arm, and then extend our uncertainty to the orthant joint space (Definition 6.4) via the combined nonlinear expectation $\mathfrak{E}$ (Definition 6.6) later. For now, we will first give a clear definition of a time consitent 'nonlinear expectation' defined on each arm. As we will focus on one particular arm, we will omit the superscript (k) for notational simplicity.

Using Assumption 6.1, we now focus on the filtered probability space $\left(\Omega, \mathbb{P}, (\mathcal{F}_t)_{0\leq t\leq T}\right)$ modelling the returns from playing a single arm. As in Peng [82], we define a nonlinear expectation as follows:

Definition 6.1. A system of operators

$$\mathcal{E}_{(t)}(\,\cdot\,) : L^\infty(\mathbb{P}, \mathcal{F}_{t+1}) \to L^\infty(\mathbb{P}, \mathcal{F}_t)$$

for $t \in \mathbb{T} := \{0, 1, ..., T-1\}$ is said to be a *single-step coherent nonlinear expectation* if it satisfies the following properties: for $X_n, X, Y \in L^\infty(\mathcal{F}_{t+1})$, with all (in)equalities holding $\mathbb{P}$-a.s, we have

(i) *Strict Monotonicity*: If $X \geq Y$ then $\mathcal{E}_{(t)}(X) \geq \mathcal{E}_{(t)}(Y)$ If, in addition, $\mathcal{E}_{(t)}(X) = \mathcal{E}_{(t)}(Y)$, then $X = Y$.

(ii) $(\mathcal{F}_t)_{t\geq 0}$-*Translation Equivariance*: $\mathcal{E}_{(t)}(X+c) = \mathcal{E}_{(t)}(X)+c$ for any for $c \in L^\infty(\mathcal{F}_t)$.

(iii) *Subadditivity*: $\mathcal{E}_{(t)}(X+Y) \leq \mathcal{E}_{(t)}(X) + \mathcal{E}_{(t)}(Y)$.

(iv) $(\mathcal{F}_t)_{t\geq 0}$-*Positive Homogeneity*: $\mathcal{E}_{(t)}(\lambda X) = \lambda \mathcal{E}_{(t)}(X)$ for any $0 \leq \lambda \in L^\infty(\mathcal{F}_t)$.

(v) *Lebesgue Property*: If $\{X_n\}_{n\in\mathbb{N}}$ is uniformly $\mathbb{P}$-a.s. bounded and $X_n \to X$ $\mathbb{P}$-a.s. then $\mathcal{E}_{(t)}(X_n) \to \mathcal{E}_{(t)}(X)$.

Remark 6.2. For simplicity, we will assume the Lebesgue property throughout this paper. In the static case, upper semi-continuity can be shown to be equivalent to the Lebesgue property over L^∞ (see [49, Corollary 4.38]). Moreover, if the operator $\mathcal{E}_{(t)}$ is induced by a BSDE (as in [31, 32, 82, 43] and many other papers), then the Lebesgue property typically follows from the L^2-continuous dependence of the BSDE on its terminal value.

Remark 6.3. It is also known (see e.g. Detlefsen and Scandolo [39]) that any coherent nonlinear expectation satisfies the $(\mathcal{F}_t)$-*regularity* property. That is, for any $X, Y \in L^\infty(\mathcal{F}_{t+1})$ and $A \in \mathcal{F}_t$,

$$\mathcal{E}_{(t)}\big(X\mathbb{I}_A + Y\mathbb{I}_{A^c}\big) = \mathbb{I}_A \mathcal{E}_{(t)}\big(X\big) + \mathbb{I}_{A^c}\mathcal{E}_{(t)}\big(Y\big).$$

In particular, $\mathcal{E}_{(t)}\big(X\mathbb{I}_A\big) = \mathbb{I}_A\mathcal{E}_{(t)}\big(X\big)$.

Definition 6.2. A system of operators

$$\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_t\big) : L^\infty(\mathbb{P},\mathcal{F}_T) \to L^\infty(\mathbb{P},\mathcal{F}_t) \quad : \quad t \in \mathbb{T}$$

is said to be an $(\mathcal{F}_t)_{t\geq 0}$-*consistent coherent nonlinear expectation* if it satisfies all properties for a single-step (Definition 6.1) with $t+1$ replaced by T and, in addition, it satisfies

(vii) $(\mathcal{F}_t)_{t\geq 0}$-*consistency*: for $X \in L^\infty(\mathcal{F}_T)$ and $0 \leq s \leq t \leq T$,

$$\mathcal{E}\big(X\big|\mathcal{F}_s\big) = \mathcal{E}\Big(\mathcal{E}\big(X\big|\mathcal{F}_t\big)\Big|\mathcal{F}_s\Big) \quad \mathbb{P}\text{-a.s..}$$

We write $\mathcal{E}(\,\cdot\,)$ for $\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_0\big)$.

The following proposition gives us a way to construct a consistent, coherent nonlinear expectation from a single-step operator.

Proposition 6.1. *There exists a natural bijection between a family of single-step coherent nonlinear expectations and consistent coherent nonlinear expectation.*

Proof. Given a system of one-step operators $\big(\mathcal{E}_{(t)}(\cdot)\big)_{t\in\mathbb{T}}$ and a system of consistent operators $\big(\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_t\big)\big)_{t\in\mathbb{T}}$. The natural bijection $\mathcal{I}$ can be given by

$$\mathcal{I}\Big(\big(\mathcal{E}_{(t)}(\cdot)\big)_{t\in\mathbb{T}}\Big) = \mathcal{E}_{(t)}\left(\mathcal{E}_{(t+1)}\left(\cdots\mathcal{E}_{(T-1)}\left(\,\cdot\,\right)\cdots\right)\right) \quad \text{and,}$$

$$\mathcal{I}^{-1}\Big(\big(\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_t\big)\big)_{t\in\mathbb{T}}\Big) = \mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_t\big)\Big|_{\mathcal{F}_{t+1}}. \qquad \square$$

Example 6.1. In Cohen [29], a method to construct a single-step nonlinear expectation (called the *DR-Expectation*) is proposed. It can be proved that, asymptotically, it corresponds to a (statistical) upper confidence bound on the expectations of given random variables. The DR-Expectation can be constructed as follows:

1. Choose a family of probability measures $\mathcal{Q}$ such that there exists a measure μ dominating $\mathcal{Q}$. This corresponds to a family of parameters in a parametric model and the dominating measure is often Lebesgue measure (or the counting measure). In this case, the Radon–Nikodym derivative is simply the usual likelihood function.

2. Choose an uncertainty aversion parameter ρ and define a family of probability measures $\tilde{\mathcal{Q}}_t$ by

$$\tilde{\mathcal{Q}}_t := \left\{\mathbb{Q} \in \mathcal{Q} \;:\; \operatorname*{ess\,sup}_{\tilde{\mathbb{Q}}\in\mathcal{Q}}\left[\log\left(\frac{d\tilde{\mathbb{Q}}}{d\mu}\bigg|_{\mathcal{F}_t}\right)\right] - \log\left(\frac{d\mathbb{Q}}{d\mu}\bigg|_{\mathcal{F}_t}\right) \leq \rho\right\}.$$

   Here, the ess sup is taken over $\mathcal{F}_t$-measurable random variables.

3. The DR-Expectation is given by

$$\mathcal{E}^{DR}_{(t)}(\,\cdot\,) := \operatorname*{ess\,sup}_{\mathbb{Q}\in\tilde{\mathcal{Q}}_t} \mathbb{E}\Big(\,\cdot\,\Big|\mathcal{F}_t\Big).$$

Remark 6.4. The DR-Expectation in [29] is defined in a more general way than here, as it is not usually assumed to satisfy positive homogeneity. This property will prove very convenient in our analysis, so we restrict to this case for this thesis.

In order to study decision making, we often require a conditional expectation defined at a stopping time. As we are working in discrete time, this is an easy construction.

Definition 6.3. Given a consistent coherent nonlinear expectation $\mathcal{E}$ and a stopping time $\tau \le T$, we define the conditional expectation at τ by

$$\mathcal{E}\Big(\,\cdot\,\Big|\mathcal{F}_\tau\Big) : L^\infty(\mathcal{F}_T) \longrightarrow L^\infty(\mathcal{F}_\tau), \qquad X \longmapsto \sum_{t=0}^{T} \mathbb{I}(\tau = t)\, \mathcal{E}\big(X\big|\mathcal{F}_t\big).$$

With this definition, the following easy observations can be made.

Proposition 6.2. *The operator $\mathcal{E}\big(\cdot\big|\mathcal{F}_\tau\big)$ satisfies the conditions of Definition 6.1 with $t+1$ replaced by T and t replaced by τ. Moreover, the family of operators $\big\{\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_\tau\big)\big\}$ satisfies the condition of Definition 6.2 with s and t replaced by stopping times σ and τ such that $0 \le \sigma \le \tau \le T$.*

Nonlinear expectations are well suited to the study of Knightian uncertainty, that is, uncertainty over the probability measure. This is most easily seen through the robust representation theorem given by Artzner et al. [7], see also Föllmer and Scheid [49] and Frittelli and Rosazza-Gianin [50]. Extensions to a dynamic setting are also considered by Detlefsen and Scandolo [39], Föllmer and Scheid [49] and Riedel [89]. We state a version of this result, which is dynamic over stopping times.

Theorem 6.3. *Let $\mathcal{E}$ be a consistent coherent nonlinear expectation. Then $\mathcal{E}$ admits the representation*

$$\mathcal{E}\Big(\,\cdot\,\Big|\mathcal{F}_\tau\Big) = \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\,\cdot\,\Big|\mathcal{F}_\tau\Big)$$

where τ is a stopping time and $\mathcal{Q} \subseteq \{\mathbb{Q} : \mathbb{Q} \approx \mathbb{P}\}$, and the essential supremum is taken in $L^\infty(\mathcal{F}_\tau, \mathbb{P})$.

Proof. See Föllmer and Schied [49, Theorem 11.22] (with further discussion in Föllmer and Penner [48]). The sensitivity assumption assumed in these references (i.e. for every nonnegative nonconstant $X \in L^\infty(\mathcal{F}_T)$, there exists $\lambda > 0$ such that $\mathcal{E}\big(\lambda X\big) > 0$) follows from strict monotonicity in our definition. □

Remark 6.5. Theorem 6.3 can be obtained by construction via considering the stability of the pasting in the family $\mathcal{Q}$. (See e.g. Bion-Nadal [17] and Artzner et al. [8]).

6.3 Uncertainty on multiple arms

We can now extend 'nonlinear expectation' to the joint space to consider the multi-armed bandit. We first define the joint space for our problem.

6.3.1 Orthant space

Definition 6.4. To model the space for the multiple arms, we define the *orthant probability space* $\left(\bar{\Omega}, \bar{\mathbb{P}}, (\mathcal{F}(s))_{s\in\mathcal{S}}\right)$ by

$$\bar{\Omega} := \prod_{k=1}^{K} \Omega^{(k)}, \quad \bar{\mathbb{P}} := \bigotimes_{k=1}^{K} \mathbb{P}^{(k)}, \quad \mathcal{F}(s) := \bigotimes_{k=1}^{K} \mathcal{F}^{(k)}_{s^{(k)}} : s = (s^{(1)}, ..., s^{(K)}) \in \mathcal{S}$$

where $\mathcal{S} := \mathbb{N}_0^K$. We call $(\mathcal{F}(s))_{s\in\mathcal{S}}$, the *orthant filtration.*

To describe a useful set of stopping times in this (multi-indexed) filtration, let

$$\mathcal{T}^{(k)} := \left\{(\mathcal{F}^{(k)}_t)_{t\geq 0}\text{-stopping times}\right\}$$

and

$$\mathfrak{T}(\mathcal{S}) := \left\{S = (S^{(1)}, ..., S^{(K)}) \quad : \quad S^{(k)} \in \mathcal{T}^{(k)}\right\}.$$

For $S \in \mathfrak{T}(\mathcal{S})$, we write

$$\mathcal{F}(S) := \bigotimes_{m=1}^{K} \mathcal{F}^{(k)}_{S^{(k)}} \quad \text{for } S \in \mathfrak{T}(\mathcal{S}).$$

Remark 6.6. Our thesis considers orthant filtration as the product of filtrations defined in different spaces. This is slightly different from Mandelbaum [77] (and the follow-up work of El Karoui and Karatzas [63]) where the orthant filtration is considered as the join of filtrations defined on the same space. This technical difference allows us to more easily define a 'nonlinear expectation' that still carries some form of independence and 'time-consistency.'

Our policies will be described by a (random) path in the space $\mathcal{S}$, which indicates how many times each arm has been played.

Definition 6.5 (Mandelbaum [77])**.** The **Mandelbaum allocation strategy** is an $\mathcal{S}$-valued random sequence $\left(\tilde{\eta}(n)\right)_{n\geq 0}$ such that

(i) $\tilde{\eta}(0) = 0$

(ii) $\tilde{\eta}(n+1) = \tilde{\eta}(n) + e^{(k)}$ for some $k \in [K] =: \{1, ..., K\}$.

(iii) $\{\tilde{\eta}(n+1) = \tilde{\eta}(n) + e^{(k)},\ \tilde{\eta}(n) = r\} \in \mathcal{F}(r)$ for all $k \in [K]$ and for all $r \in \mathcal{S}$.

Here, $e^{(k)}$ denotes the kth unit vector in $\mathcal{S}$.

Remark 6.7. A Mandelbaum allocation strategy $\big(\tilde{\eta}(n)\big)_{n\geq 0}$ can also be represented by its increments, in particular, by a sequence of decision variables $(\rho_n)_{n\geq 0}$ taking values in $[K]$ (as a random decision (A_t) in section 1.1.1). In other words, we can define $(\rho_n)_{n\geq 0}$ such that $\{\rho_n = k\} = \{\tilde{\eta}(n+1) = \tilde{\eta}(n) + e^{(k)}\}$. We can write down the expected total discounted cost under $\bar{\mathbb{P}}$ by

$$\inf_{\tilde{\eta}} \mathbb{E}^{\bar{\mathbb{P}}}\Big(\sum_{n=1}^{\infty} \beta^n h^{(\rho_{n-1})}(t_n^{\rho})\Big) \quad \text{where} \quad t_n^{\rho} := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1}). \tag{6.3.1}$$

6.3.2 Independence under Nonlinear Expectation

In the classical Gittins' theorem, 'independence' is crucial to separate the behaviour of different arms. In the robust representation (Theorem 6.3) we have seen that a nonlinear expectation can be viewed as the supremum of classical expectations over a family of probability measures. Therefore, the notion of independence between arms becomes ambiguous, as statistical independence is based on the probability measure. Thanks to our explicit construction of the space (Definition 6.4), we can explicitly construct a nonlinear expectation space where each arm remains 'independent'.

Remark 6.8. In [83], Peng proposed a definition of independence for a nonlinear expectation. In his approach, independence is not a symmetric relation but typically describes independence based on the order of events: often 'Y is independent of X' when Y occurs after X. In the setting of multiple arms, the order of events cannot be pre-identified, as it depends on the control chosen. Hence, it is not clear how to exploit the independence notion of [83] in this setting.

Let us make the last universal assumption in our paper, which describes model uncertainty for each individual arm in our problem.

Assumption 6.3. *For each $k \in [K] = \{1, ..., K\}$, we have an $(\mathcal{F}_t^{(k)})_{t\geq 0}$-consistent coherent nonlinear expectation, $(\mathcal{E}^{(k)}(\,\cdot\,|\mathcal{F}_\tau))_{\tau\in\mathcal{T}^{(k)}}$, defined on the space $L^\infty(\Omega^{(k)}, \mathbb{P}^{(k)})$.*

By Theorem 6.3, $\mathcal{E}^{(k)}$ admits the representation

$$\mathcal{E}^{(k)}\Big(\,\cdot\,\Big|\mathcal{F}_{S^{(k)}}^{(k)}\Big) = \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}^{(k)}} \mathbb{E}^{\mathbb{Q}}\Big(\,\cdot\,\Big|\mathcal{F}_{S^{(k)}}^{(k)}\Big)$$

whenever $S^{(k)}$ is an $(\mathcal{F}_t^{(k)})$-stopping time.

Definition 6.6. We define the *partially consistent orthant nonlinear expectation* $(\mathfrak{E}_S)_{S\in\mathfrak{T}(\mathcal{S})}$, to be the family of operators

$$\begin{aligned} \mathfrak{E}_S : L^\infty\big(\bar{\Omega}, \bar{\mathbb{P}}, \mathcal{F}(T)\big) &\longrightarrow L^\infty\big(\bar{\Omega}, \bar{\mathbb{P}}, \mathcal{F}(S)\big) \\ X &\longmapsto \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\big(X\big|\mathcal{F}(S)\big) \end{aligned}$$

where, with $\mathcal{Q}^{(k)}$ as in Assumption 6.3,

$$\mathcal{Q} := \left\{ \bigotimes_{m=1}^{K} \mathbb{Q}^{(k)} \ \text{ for } \ \mathbb{Q}^{(k)} \in \mathcal{Q}^{(k)} \right\}.$$

We also write $\mathfrak{E}$ for $\mathfrak{E}_0$.

Remark 6.9. As $\bar{\mathbb{P}} = \bigotimes_{k=1}^{K} \mathbb{P}^{(k)}$ is a dominating measure for $\mathcal{Q}$, we easily observe that if $X = Y$ $\bar{\mathbb{P}}$-a.s., then $\mathfrak{E}_S(X) = \mathfrak{E}_S(Y)$ $\bar{\mathbb{P}}$-a.s. for all $S \in \mathfrak{T}(\mathcal{S})$.

Proposition 6.4. *The system of operators $(\mathfrak{E}_S)$ satisfies the following properties.*

(i) All the properties in Definition 6.1 (with $\mathcal{E}_{(t)}$ replaced by $\mathfrak{E}_S$ and $\mathcal{F}_t$ replaced by $\mathcal{F}(S)$) hold for the operator $\mathfrak{E}_S$: $S \in \mathfrak{T}(\mathcal{S})$ (i.e. strict monotonicity, translation equivariance, subadditivity, positive homogeneity and the Lebesgue Property).

(ii) Sub-consistency: For $S, S' \in \mathfrak{T}(\mathcal{S})$ with $S \leq S'$, we have

$$\mathfrak{E}_S(\,\cdot\,) \leq \mathfrak{E}_S\left(\mathfrak{E}_{S'}(\,\cdot\,)\right) \quad \bar{\mathbb{P}}\text{-a.s.}$$

In particular, for any measurable X, if $\mathfrak{E}_{S'}(X) \leq 0$ $\bar{\mathbb{P}}$-a.s., then for any $A \in \mathcal{F}_{S'}$ we have $\mathfrak{E}_S(\mathbb{I}_A X) \leq 0$ $\bar{\mathbb{P}}$-a.s.

(iii) Independence: Let Y be a random variable on $(\bar{\Omega}, \mathcal{F}(T))$ given by

$$Y(\omega^{(1)}, \ldots, \omega^{(k)}) = X^{(1)}(\omega^{(1)}) \times \cdots \times X^{(k)}(\omega^{(k)}) \qquad \bar{\mathbb{P}}\text{-a.s.},$$

where, for each $k \in [K]$, we have a non-negative random variable $X^{(k)}$ defined on $(\Omega^{(k)}, \mathcal{F}^{(k)})$. Then

$$\mathfrak{E}_S(Y) = \mathcal{E}^{(1)}\left(X^{(1)}\middle|\mathcal{F}^{(1)}_{S^{(1)}}\right) \times \cdots \times \mathcal{E}^{(k)}\left(X^{(K)}\middle|\mathcal{F}^{(K)}_{S^{(K)}}\right) \quad \bar{\mathbb{P}}\text{-a.s.}$$

(iv) Marginal projection: For a given $m \in \mathcal{K}$, let X be a random variable defined on $(\Omega^{(k)}, \mathcal{F}^{(k)}_{T^{(k)}})$. Define $\tilde{X} : \bar{\Omega} \to \mathbb{R}$ by $\tilde{X}(\omega^{(1)}, \ldots, \omega^{(K)}) = X(\omega^{(k)})$. We then have

$$\mathfrak{E}_S(\tilde{X}) = \mathcal{E}^{(k)}\left(X\middle|\mathcal{F}^{(k)}_{S^{(k)}}\right) \quad \bar{\mathbb{P}}\text{-a.s.}$$

In the proposition above, we have seen that $\mathfrak{E}$ is sub-consistent on the orthant filtration. However, $\mathfrak{E}$ is *not* consistent in the sense of Definition 6.2, i.e. if $S \leq S'$ (componentwise), it is not necessarily the case that $\mathfrak{E}_S(\,\cdot\,) = \mathfrak{E}_S\left(\mathfrak{E}_{S'}(\,\cdot\,)\right)$. A counterexample can be easily constructed based on the following:

Example 6.2. Let X and $\tilde{X}$ be random variables taking values in $\{0, 1\}$ and defined on different spaces Ω and $\tilde{\Omega}$. Let $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ be families of probability measures defined on these spaces. Suppose that for all $p \in [0, 1]$ there exists $\mathbb{Q} \in \mathcal{Q}$ such that $\mathbb{Q}(X = 0) = p$ and that for all $\tilde{\mathbb{Q}} \in \tilde{\mathcal{Q}}$, $\tilde{\mathbb{Q}}(\tilde{X} = 0) = 1/2$. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a given function. Then it is easy to show that

$$\begin{aligned}\sup_{\mathbb{Q},\tilde{\mathbb{Q}}} \mathbb{E}^{\mathbb{Q}\otimes\tilde{\mathbb{Q}}}\big(f(X, \tilde{X})\big) &= \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\sup_{\tilde{\mathbb{Q}}} \mathbb{E}^{\tilde{\mathbb{Q}}}\big(f(x, \tilde{X})\big)\Big|_{x=X}\Big) \\ &= \max\Big(\frac{f(0,0) + f(0,1)}{2}, \frac{f(1,0) + f(1,1)}{2}\Big)\end{aligned}$$

but

$$\sup_{\tilde{\mathbb{Q}}} \mathbb{E}^{\tilde{\mathbb{Q}}}\Big(\sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}\big(f(X,\tilde{x})\big)\Big|_{\tilde{x}=\tilde{X}}\Big) = \frac{\max\{f(0,0),f(1,0)\}}{2} + \frac{\max\{f(0,1),f(1,1)\}}{2}.$$

By considering $\mathcal{F}_1^{(1)} = \sigma(X)$ and $\mathcal{F}_1^{(2)} = \sigma(\tilde{X})$, and defining nonlinear expectation using supremum over the family $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$, the above result shows that the joint operator $\mathfrak{E}$ is not consistent. In particular, we can find a function f such that

$$\mathfrak{E}\Big(\mathfrak{E}_{(0,1)}\Big(f(X,\tilde{X})\Big)\Big) \neq \mathfrak{E}\Big(\mathfrak{E}_{(1,0)}\Big(f(X,\tilde{X})\Big)\Big).$$

6.4 Optimality Criteria

The Gittins index theorem states that an index-based strategy can give an optimal solution to (6.3.1); where an option with the minimum index is played at each time. The index of each option can be evaluated separately. Unfortunately, Gittins index theorem does not extend to the robust case when forcing time consistency through the robust Bellman equation.

6.4.1 On the Robust Gittins Index

Under a Markovian assumption, Caro and Gupta [23] consider 'robust' Gittins index based on the Robust Bellman equation studied in Iyergar [60] and Nilim and El Ghaoui [80]. (Kim and Lim [66] considered a similar work with an additional penalty in the formulation.)

The following assumptions are used in Caro and Gupta [23] (translated into our notation):

(i) The cost process $(h^{(k)}(t))$ is driven by some underlying finite-state predictable process $(X_t^{(k)})_{t\geq 0}$ taking values in $\mathcal{X}^{(k)}$. i.e. we have $h^{(k)}(t) = \tilde{h}^{(k)}(X_t^{(k)})$ for some deterministic function $\tilde{h}^{(k)}$.

(ii) Ambiguity is described by families of transition matrices, $(\mathcal{U}^{(k)})_{m\in\mathcal{M}}$ for the dynamics of $X^{(k)}$.

The construction of their robust index is then based on Whittle indexability [111]. In particular, they reduce the problem to considering two arms, where one arm always generates a constant cost γ and the other arm is identical to the kth arm. The worst-case expected cost obtained when starting in state i in the kth arm, $V^{(k)}(i)$, allowing any combination of transition rates, will then satisfy the robust dynamic programming principle, that is,

$$V^{(k)}(i) = \min\left(\tilde{h}^{(k)}(i) + \beta \sup_{P\in\mathcal{U}^{(k)}} \sum_{j\in\mathcal{X}^{(k)}} P_{ij}V^{(k)}(j), \quad \frac{\gamma}{1-\beta}\right), \qquad i \in \mathcal{X}^{(k)}. \tag{6.4.1}$$

Caro and Gupta show that (6.4.1) is Whittle indexable (as discussed in section 2.1.3), i.e. if $D^{(k)}(\gamma) \subseteq \mathcal{X}^{(k)}$ is the set of states for which it is optimal to rest the kth

arm when the reward of the constant arm is γ, then $D^{(k)}(\gamma)$ increases monotonically from ϕ to $\mathcal{X}^{(k)}$ as γ increases from $-\infty$ to $+\infty$. The index of the kth arm at state i is the unique value γ such that the player is indifferent between playing the kth arm and the constant arm.

This index can be characterised by

$$\tilde{\gamma}^{(m)}(i) = \inf_{\tau \in \mathcal{T}^{(m)}, \tau \geq 1} \sup_{\mathbb{Q} \in \mathcal{Q}^{(m)}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=0}^{\tau-1} \beta^t \tilde{h}^{(m)}(X_t^{(m)}) \middle| X_0^{(m)} = i \right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=0}^{\tau-1} \beta^t \middle| X_0^{(m)} = i \right)}. \tag{6.4.2}$$

where $\mathcal{Q}^{(m)}$ is the family of measures corresponding to the family of transition matrices $\mathcal{U}^{(m)}$.

Unfortunately, as discussed in Caro and Gupta [23], the robust Gittins index (6.4.2) does not yield a strategy optimising the robust Bellman equation:

$$V(i_1, ..., i_K) = \min_{k \in [K]} \left(\tilde{h}^{(k)}(i_k) + \beta \sup_{P \in \mathcal{U}^{(k)}} \sum_{j \in \mathcal{X}^{(k)}} P_{i_k j} V(i_1, ..., i_{k-1}, j, i_{k+1}, ..., i_K) \right), \ i_k \in \mathcal{X}^{(k)}.$$

In short, this non-optimality arises due to the fact that we force time-consistency through the robust Bellman equation. Even if we assume at first that the arms do not depend on each other, the robust Bellman equation introduces dependency between arms. In particular, at equilibrium, the adversary (who determines the transition probabilities for each arm) may choose differently depending on the state of *all* arms, rather than just the arm of interest. This means that the arms introduce a form of 'independence' and therefore that Gittins' theorem fails. This is also counterintuitive if we begin our modelling by assuming the evolution of arms do not depend on each other.

6.4.2 Optimality Criteria

In order to understand what sense of optimality the robust index strategy *does* satisfy, we will first consider a form of optimality criteria that El Karoui and Karatzas [63] used to prove the classical Gittins' theorem. We then propose a new form of optimality in terms of compensators of the value function. This can be seen as a relaxation of the dynamic programming principle to address a control problem under an inconsistent nonlinear operator. We will show later in chapter 7 and 8 that the strategy given by robust Gittins index satisfies this optimality criterion.

Our result not only proves the optimality of the robust index of Caro and Gupta [23] but also extends to a more general setting. The novelty of the general extension is to allow the cost to be continuous-valued and non-Markovian. This enables the study of various numerical methods to estimate our probabilistic state in the learning problem. In contrast, the numerical method in Caro and Gupta [23] is limited to the finite state Markov process. A simple numerical example of the robust index shows some qualitative peculiarities given the interaction between uncertainty aversion and learning. We will discuss this in section 7.3 of chapter 7.

Let us consider an abstract stochastic control problem on a space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ in which a choice of control ρ results in an instantaneous cost process $\left(g^{\rho}(n)\right)_{n \geq 1}$.

We may view $g^\rho(n)$ as a cost occured at time n. For example, we have $g^\rho(n) := \beta^n h^{(\rho_{n-1})}(t_n^\rho)$ in (6.3.1). We can also define the filtration of information obtained up to time n when following ρ by

$$\mathcal{G}_n^\rho := \Big\{ A \in \mathcal{F}(T) \ : \ A \cap \{\tilde{\eta}(n) = r\} \in \mathcal{F}(r) \quad \forall r \in \mathcal{S} \Big\}, \tag{6.4.3}$$

where $\tilde{\eta}$ is the corresponding Mandelbaum allocation sequence (Definition 6.5, Remark 6.7). We will discuss this filtration in detail in Remarks 7.5 and 7.6.

Remark 6.10. It is clear from the definition that the strategy process $(\rho_n)_{n\geq 0}$ is $(\mathcal{G}_n^\rho)_{n\geq 0}$-adapted. We will show later that the cost process $g^\rho(n) := \beta^n h^{(\rho_{n-1})}(t_n^\rho)$ (as in (6.3.1)) is also adapted with respect to $\mathcal{G}_n^\rho$.

Suppose that we are given a nonlinear expectation operator $\mathfrak{E}$, as in Definition 6.6, and consider a minimization problem over the space of Mandelbaum allocation strategies, as represented by their equivalent form ρ (Remark 6.7). The process ρ not only describes our strategy and the corresponding cost, but also generates the observed filtration. Therefore, at any point in time, it does not make sense to compare strategies unless those strategies yield the same stage of information at the considered time.

Definition 6.7. We say a strategy ρ and ρ' are **strategically equivalent** at time N, denoted by $\rho \sim_N \rho'$, if $\rho_n = \rho'_n$ for all $n \leq N$.

Remark 6.11. For every strategy ρ, we have $\rho \sim_0 \rho^*$.

We can now give a standard form of optimality which is often considered in the literature when we have a consistent nonlinear expectation operator.

Definition 6.8. We say a strategy ρ^* is a *strong optimum* if for every strategy ρ such that $\rho \sim_N \rho^*$, we have

$$\mathfrak{E}\left(\mathbb{I}_A\left(\sum_{n=N+1}^{\infty} g^{\rho^*}(n)\right)\right) \leq \mathfrak{E}\left(\mathbb{I}_A\left(\sum_{n=N+1}^{\infty} g^{\rho}(n)\right)\right) \qquad \text{for all } A \in \mathcal{G}_N^\rho (= \mathcal{G}_N^{\rho^*}).$$

Remark 6.12. When $\mathfrak{E}$ is replaced by an $(\mathcal{F}_n)$-consistent nonlinear expectation (and $(\mathcal{G}_N^\rho)$ is replaced by $(\mathcal{F}_N)$), the strong optimality is simply

$$\mathcal{E}\left(\sum_{n=N+1}^{\infty} g^{\rho^*}(n)\middle|\mathcal{F}_N\right) = \operatorname*{ess\,inf}_{\rho\sim_N\rho^*} \mathcal{E}\left(\sum_{n=N+1}^{\infty} g^{\rho}(n)\middle|\mathcal{F}_N\right).$$

A standard approach to tackle the decision making under time-inconsistency (nonlinear expectation) operator is to define 'the optimal strategy' through the solution of the robust Bellman equation ([60, 80]) which is what considered in Caro and Gupta [23]. Using the tower property, we can show that the strong optimum under $(\mathcal{F}_n)$-consistent nonlinear expectation is equivalent to the solution to the robust Bellman equation.

In the bandit setting, our considered operator is not necessary (time-)consistent wheresas Gittins index is defined through a consistent operator(i.e. we can write

(6.4.1) in the form (7.1.1)). Hence, it does not seem natural that Gittins index would satisfy the artificial notion of 'the optimal (consistent) strategy' through the Robust Bellman equation where time-consistency is forced.

To allow ourselves to prove an optimality of Gittins index strategy under an inconsistent operator, we propose an alternative notion of optimality which is inspired by the martingale optimality.

6.4.3 C-Optimality

Let consider an $(\mathcal{F}_n)$-consistent nonlinear expected operator $\mathcal{E}$. Suppose that we wish to solve the minimization problem

$$V_N = \operatorname*{ess\,inf}_{\rho} \mathcal{E}\left(\sum_{n=N+1}^{\infty} g^{\rho}(n) \middle| \mathcal{F}_N \right)$$

For a given strategy ρ, we define a process $X_N^{\rho} := \sum_{n=1}^{N} g^{\rho}(n) + V_N$. Under a mild condition, we know from the martingale optimality principle that (X_N^{ρ}) is a submartingale for every strategy ρ and it is a martingale for an optimal strategy ρ^*.

By using the Doob–Meyer decomposition for nonlinear expectation (see e.g. Cohen [30]), we can write

$$X_N^{\rho} := M_N^{\rho} + \sum_{n=1}^{N} C^{\rho}(n)$$

where (M_N^{ρ}) is a martingale and $(C^{\rho}(n))$ is a non-negative predictable process with $C^{\rho}(n) \equiv 0$ for the optimal strategy ρ^*.

By rearranging the equation above, we can show that

$$\mathcal{E}\left(\sum_{n=N+1}^{\infty} \Big(g^{\rho}(n) - C^{\rho}(n) \Big) \middle| \mathcal{F}_N \right) = -V_N.$$

Moreover, for an optimal strategy ρ^*, we have

$$\sum_{n=N+1}^{L} C^{\rho^*}(n) \le \sum_{n=N+1}^{L} C^{\rho}(n) \qquad \text{for all } N, L.$$

Inspired by the analysis above, we propose to consider an alternative notion of the optimality.

Definition 6.9. We say a strategy ρ^* is *C-optimal* if there exists a $(\mathcal{G}_n^{\rho^*})$-adapted process (V_n) (called a *value process*) and a collection of random variables $(C_N^{\rho}(n))_{N, n \ge N+1, \rho \sim_N \rho^*}$ (called a (sub-)compensator) such that

(i) $n \mapsto C_N^{\rho}(n)$ is an $(\mathcal{G}_n^{\rho})$-predictable process,

(ii) $N \mapsto C_N^{\rho}(n)$ is non-increasing,

(iii) For every strategy $\rho \sim_N \rho^*$,

$$\mathfrak{E}\Big(\mathbb{I}_A \sum_{n=N+1}^{\infty} \Big(g^{\rho}(n) - C_N^{\rho}(n)\Big)\Big) \geq \mathfrak{E}\Big(\mathbb{I}_A V_N\Big) \qquad \text{for all} \quad A \in \mathcal{G}_N^{\rho} \tag{6.4.4}$$

with equality holds for ρ^*,

(iv) For every strategy $\rho \sim_N \rho^*$,

$$\sum_{n=N+1}^{L} C_N^{\rho^*}(n) \leq \sum_{n=N+1}^{L} C_N^{\rho}(n) \qquad \text{for all } L \geq N+1. \tag{6.4.5}$$

We can see $\big(C^{\rho}(n)\big)$ as a process to compensate the cost, in order to obtain our referencing value function. This approach is loosely related to the capital requirement approach discussed by Frittelli and Scandolo [51]. We can interpret Definition 6.9 as requiring that the (sub-)compensators $(C_N^{\rho}(n))$

(i) is known one-step in advanced before observing the cost (i).

(ii) consistently (sub-)compensate the cost. In particular, as the time elapse, we obtain more information and thus require the same amount or possibly smaller amount to (sub-)compensate (ii).

(iii) complement the extra cost occurred for a sub-optimal strategy (iii).

(iv) are bounded below by a compensator of a particular strategy ρ^* which we call 'optimal' (iv).

Remark 6.13. We have mentioned the robust Bellman equation ([60, 80]) as an approach to force time-consistency in our decision making. The fundamental idea of this approach is to freeze our value function and propagate its value backward in time. In particular, suppose we have V_{n+1}^* as our expected remaining cost at time n. We then define an optimal strategy at time n to be a strategy ρ_n such that $g^{\rho}(n) + V_{n+1}^*$ is optimized.

Another approach to ensure time-consistency was proposed by Strotz [102] and Pollak [85] and developed further in Peleg and Yaari [81] and Koopmans [69]. Recent extensions include Björk and Murgoci [19], Björk, Khapko and Murgoci [18], Yong [113] and Hu, Jin and Zhou [58]. The essential idea of this approach is to use backward induction, that is, to suppose that we consider the problem with horizon L and the optimal control is determined after time n, that is, $(\rho_{n+1}^*, ..., \rho_L^*)$ is known; we then find a control ρ_n^* at time n to optimize over the space of possible strategies $\{\rho \ : \ (\rho_{n+1}, ..., \rho_L) = (\rho_{n+1}^*, ..., \rho_L^*)\}$. This idea is then extended by searching for Nash equilibria, to allow for non-uniqueness of the optimal controls.

We have discussed earlier that the robust Bellman approach may introduce some form of non-required dependency in our system. Hence, Gittins index strategy is not optimal under this approach of the optimality. On the other hand, when considering a system of bandits, the measurability of our future states are determined by our

current action. Therefore, the σ-algebra that is used to define the future control $(\rho_k^*)_{k\geq n+1}$ would restrict our choice of our current control. This means that we cannot directly consider the Strotz–Pollak approach for the bandit setting as we cannot freeze our future control without freezing our current control.

6.4.4 Endowment Effect

One of a natural question to ask is can we give an interpretation of the C-optimality (Definition 6.9) in terms of a classical strong optimality (Definition 6.8). To see this, we will consider an endowment effect through the strong optimality.

Example 6.3. Let H and G be random variables representing the cost of two strategies and $\mathcal{Q}$ be a family of probability measures such that H and G are independent under each $\mathbb{Q} \in \mathcal{Q}$. Suppose $\{\mathbb{E}^{\mathbb{Q}}(H)\}_{\mathbb{Q}\in\mathcal{Q}} = [\underline{h}, \bar{h}]$ and similarly for G. Suppose further that $\bar{h} < \bar{g}$ but $\bar{h} - \underline{g} > \bar{g} - \underline{h}$. Then for $\mathfrak{E}(\,\cdot\,) := \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}(\,\cdot\,)$, we have

$$\mathfrak{E}(H) < \mathfrak{E}(G) \quad \text{but} \quad \mathfrak{E}\Big(H - \frac{H+G}{2}\Big) > \mathfrak{E}\Big(G - \frac{H+G}{2}\Big). \tag{6.4.6}$$

One may see that, without any endowment, we strictly prefer H to G whereas our preference reverses with an endowment $(H+G)/2$. We know that in the classical linear expectation theory (where the classical Gittins theorem holds), an endowment does not affect our preference in the strategy.

In this section, we will show that C-optimality is nearly equivalent to a strong optimality up to an endowment when our nonlinear expectation is time-consistent.

The following proposition follows from the definition of C-optimality and monotonicity of nonlinear expectation.

Proposition 6.5. *Let ρ^* be a C-optimal strategy with a predictable compensator $(C_N^{\rho^*}(n))$. Then for every $\rho \sim_N \rho^*$ and $A \in \mathcal{G}_N^{\rho^*}$,*

$$\mathfrak{E}\bigg(\mathbb{I}_A \sum_{n=N+1}^{\infty} \big(g^{\rho^*}(n) - C_N^{\rho^*}(n)\big)\bigg) \leq \mathfrak{E}\bigg(\mathbb{I}_A \sum_{n=N+1}^{\infty} \big(g^{\rho}(n) - C_N^{\rho^*}(n)\big)\bigg). \tag{6.4.7}$$

Let consider the case when $\mathcal{G}_N^{\rho} = \mathcal{F}_N$ for every strategy ρ and pretend that $\mathfrak{E}$ is an $(\mathcal{F}_n)$-consistent nonlinear expectation operator. Then (6.4.7) says that the C-optimality implies strong optimality with a predictable endowment. We will now show that the reverse also holds when our operator is consistent.

Definition 6.10. Let $\mathcal{E}$ be an $(\mathcal{F}_n)$-consistent nonlinear expectation. We say a strategy ρ^* is *optimal up to a predictable endowment* if there exists a family of random variables $(D_N(n))$ such that

(i) $n \mapsto D_N(n)$ is an $(\mathcal{F}_n)$-predictable process,

(ii) $N \mapsto D_N(n)$ is non-increasing,

(iii) For every strategy $\rho \sim_N \rho^*$, for all $A \in \mathcal{G}_N^\rho$,

$$\mathcal{E}\left(\mathbb{I}_A \sum_{n=N+1}^{\infty} \left(g^{\rho^*}(n) - D_N(n)\right)\right) \leq \mathcal{E}\left(\mathbb{I}_A \sum_{n=N+1}^{\infty} \left(g^{\rho^*}(n) - D_N(n)\right)\right) \tag{6.4.8}$$

or equivalently,

$$\mathcal{E}\left(\sum_{n=N+1}^{\infty} \left(g^{\rho^*}(n) - D_N(n)\right)\middle|\mathcal{F}_N\right) \leq \mathcal{E}\left(\sum_{n=N+1}^{\infty} \left(g^{\rho^*}(n) - D_N(n)\right)\middle|\mathcal{F}_N\right)$$

Proposition 6.6. *Suppose that ρ^* is an optimal strategy up to a predictable endowment, then ρ^* is C-optimal.*

Proof. Take $C_N^\rho(n) = D_N(n)$ and $V_N = \mathcal{E}\left(\sum_{n=N+1}^{\infty}\left(g^{\rho^*}(n) - D_N(n)\right)\middle|\mathcal{F}_N\right)$. □

We can see from Proposition 6.5 and Proposition 6.6 that when we have a time-consistent operator C-optimal is equivalent to the strong optimality up to predictable endowment. In the upcoming chapters we will show that Gittins theorem holds under C-optimality criteria under an operator $\mathfrak{E}$. This means that we prove that Gittins theorem is (strong) optimal upto some predictable endowment.

Remark 6.14. It is an open question that under which condition that the C-optimum point is unique. In the most trivial case when our operator $\mathfrak{E}$ is simply a classical expectation, the endowment never affect our evaluation and thus it is reduced to the uniqueness of the optimal solution in the classical setting.

Chapter 7

Gittins' theorem under Uncertainty Aversion

In the previous chapter, we introduced nonlinear expectations as a tool to model uncertainty aversion over multiple filtrations. We also presented a form of (dynamic) C-optimality, which can be seen as an approach to overcome time-inconsistency.

This chapter will introduce a formal definition of the allocation strategy, which allows marginal projection in our analysis. We then consider those strategies and outline the proof for the robust Gittins' theorem under (dynamic) C-optimality. At the end of this chapter, we will discuss a simple numerical example suggesting some connections to behavioural finance. The full proof of the robust Gittins' theorem will be deferred to Chapter 8.

This chapter is based on the paper [35].

7.1 Robust Gittins theorem

Let us recall that the objective of our problem is to dynamically allocate a single resource amongst K arms to minimise the total discounted cost. We have made a few assumptions to model uncertainty in the cost process, which can be founded in Assumptions 6.1, 6.2 and 6.3.

We also introduce a Mandelbaum allocation strategy (Definition 6.5) and the equivalent notion ρ (Remark 6.7) representing the choice of our control. We are now ready to establish a robust Gittins' theorem with optimality in the sense of Definition 6.9. Our robust Gittins' theorem generalises the result of El Karoui and Karatzas [63] to the uncertain case (via Proposition **??**(ii)). One may also see this result as providing a sense of optimality for the index strategy considered by Caro and Gupta [23] and Li [74].

7.1.1 Robust Gittins' theorem

We will first give an alternative definition to the robust Gittins' index inspired by Weber [108], which is more convenient to use in our analysis.

Definition 7.1. For each $s \geq 0$, we define the *robust Gittins index* of the mth arm by

$$\gamma^{(k)}(s) := \operatorname{ess\,inf}\left\{\gamma : \operatorname*{ess\,inf}_{\tau \in \mathcal{T}^{(k)}(s)} \mathcal{E}^{(k)}\left(\sum_{t=1}^{\tau} \beta^t \left(h^{(k)}(s+t) - \gamma\right) \,\middle|\, \mathcal{F}_s^{(k)}\right) \leq 0\right\} \tag{7.1.1}$$

where $\mathcal{T}^{(k)}(s)$ is the space of positive $(\mathcal{F}_{s+t}^{(k)})_{t\geq 0}$-stopping times[1] and the outer essential infimum is taken in $L^\infty(\mathcal{F}_s^{(k)})$.

By using the results proved in the Chapter 8, we can write the robust Gittins index explicitly. We present this result here to clarify that our results agree with other approaches [23, 54, 63] as discussed in Chapter 6, but make no use of it in subsequent arguments.

Theorem 7.1. *Let $\mathcal{Q}^{(k)}$ be the family of probability measures defined in Theorem 6.3.*

$$\gamma^{(k)}(s) = \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}^{(k)}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau} \beta^t h^{(k)}(s+t) \middle| \mathcal{F}_s^{(k)}\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau} \beta^t \middle| \mathcal{F}_s^{(k)}\right)}$$[2]

Recall the partially consistent orthant nonlinear expectation $\mathfrak{E}$ induced by the family $(\mathcal{E}^{(k)})_{m\in[K]}$ as given in Definition 6.6. We can obtain an optimal allocation strategy by considering the following theorem.

Theorem 7.2 (Robust Gittins theorem)**.** *Suppose that for each $m \in \mathcal{M}$. Let $\psi_n^{(m)}$ be the total number of trials of the mth bandit before the nth play of the system. i.e. $\psi_n^{(m)} := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k^* = m)$ (given an allocation strategy ρ^* up to time $n-1$).*

Then the allocation strategy ρ^* given (recursively) by

$$\rho_n^* := \min\left\{m \in \mathcal{M} \;:\; m \in \arg\min_k \gamma^{(k)}(\psi_n^{(k)})\right\}$$

is C-optimal (Definition 6.9) under $\mathfrak{E}$ for the cost

$$g^\rho(n) = \beta^n h^{(\rho_{n-1})}(t_n^\rho) \quad \text{where} \quad t_n^\rho = \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1}).$$

Remark 7.1. We choose ρ^* to be the minimum value in the (random) set of arms with minimum Gittins index $\{\arg\min_k \gamma^{(k)}(\psi_n^{(k)})\}$. This is a simple method of symmetry breaking, in order to avoid complexities due to measurable selection. In fact, any choice of $\rho_n^* \in \{\arg\min_k \gamma^{(k)}(\psi_n^{(k)})\}$ also yields C-optimality.

Remark 7.2. The robust Gittins' theorem states that an optimal choice is given by always playing an arm with the lowest robust Gittins index. At each time, the indices of unplayed arms do not change. This leads to a form of consistency in the values associated with different arms, resulting in consistency in the decision making, even though $\mathfrak{E}$ is not consistent.

[1] Equivalently, for $\tau \in \mathcal{T}(s)$, $s+\tau$ is an $(\mathcal{F}_t)_{t\geq 0}$-stopping time.

[2] This robust index is similar to (6.4.2) but there is some difference due to the difference between adaptability and predicatability (See Remark 2.2)

7.2 Sketch proof for the Robust Gittins' theorem

We will separate the proof into two parts: In Part A, we analyse a single-armed bandit in a robust setting. In Part B, we combine K arms together. We summarise the structure and approach of the proof here. The main body of the rigorous proof can be found in section 8.1 and 8.2, respectively.

7.2.1 Single-armed bandit optimality

We begin by considering the play of the kth arm (with the superscript (k) omitted).

Step A.1 Observe that the robust Gittins index is the *minimum* compensation for which we are willing to continue to play the arm (with compensation).

By minimality, the net expected cost under optimal play must be zero (Theorem 8.1), i.e.

$$\operatorname*{ess\,sup}_{\tau \in \mathcal{T}(s)} \mathcal{E}\left(\sum_{t=1}^{\tau} \beta^t \left(h(s+t) - \gamma(s) \right) \;\middle|\; \mathcal{F}_s \right) = 0.$$

In particular, we must have a positive net expected cost when continuing to play to an arbitrary stopping time with the robust Gittins index as a compensating reward, i.e. starting at time s, for any subsequent stopping time $\tau \in \mathcal{T}(s)$, we have

$$\mathcal{E}\left(\sum_{t=1}^{\tau} \beta^t \left(h(s+t) - \gamma(s) \right) \;\middle|\; \mathcal{F}_s \right) \geq 0. \tag{7.2.1}$$

Step A.2 We view the process γ as the 'average' cost of playing the arm. Once the process $(\gamma(t))_{t \geq s}$ exceeds $\gamma(s)$, the reward $\gamma(s)$ will no longer be sufficient to encourage continued play; so it will be optimal to stop. In particular, the stopping time

$$\sigma(s, \gamma(s)) := \inf\{\theta \geq 1 : \gamma(s+\theta) > \gamma(s)\} \tag{7.2.2}$$

yields equality in (7.2.1) (Theorem 8.6).

Step A.3 Imagine that, whenever the arm (with compensating reward) is no longer attractive to play, we were to increase the compensation sufficiently to make ourselves indifferent to continuing. The expected value of future loss, with this increased compensation, must again be zero (Proposition 8.7). The offered compensation can be written as a running maximum of the robust Gittins index process and we can express the expected return

$$\mathcal{E}\left(\sum_{t=1}^{\infty} \beta^t \big(h(t) - \Gamma(t) \big) \right) = 0 \qquad \text{where} \quad \Gamma(t) := \max_{0 \leq \theta \leq t-1} \gamma(\theta). \tag{7.2.3}$$

With the compensating reward $(\Gamma(t))$, we are always willing to continue to play. In particular, at any point in time, we have a non-positive expected future cost (Theorem

8.9), i.e.

$$\mathcal{E}\left(\sum_{t=N+1}^{\infty} \beta^t \left(h(t) - \Gamma(t)\right) \middle| \mathcal{F}_N \right) \leq 0 \quad \text{for all} \quad N = 0, 1, \ldots \tag{7.2.4}$$

Step A.4 Now suppose we were to take a break from playing for some period, and then resume our earlier strategy. In this case, we may lose some expected profit (Equation (7.2.4)) due to the discount effect of the delay. .

By (7.2.3), the total reward of this game is zero. Therefore, the delay of getting the reward must result in a possibly worse outcome. In Theorem 8.10, we use this observation, together with the robust representation theorem (Theorem 6.3), to show that for any fixed $\epsilon > 0$ there is a probability measure $\mathbb{Q} \in \mathcal{Q}$ such that, for every decreasing predictable process $(\alpha(t))$ taking values in $[0, 1]$,

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\infty} \alpha(t)\beta^t\big(h(t) - \Gamma(t)\big) \right) \geq -\epsilon. \tag{7.2.5}$$

Remark 7.3. Step A.4 is the key point in which positive homogeneity of $\mathcal{E}$ is used. A predictable process $(\alpha(t))$ represents the delay due to taking a break to play another arm. In step A.3, we choose the compensator such that the total expected return is zero but the arm is always attractive to be played. (i.e. we always have a reward for the future.) We therefore cannot expect a better outcome than zero if we delay our play. Mathematically, one can replace positive homogeneity and subadditivity by convexity and the property that: if $\mathcal{E}(X|\mathcal{F}_t) \leq 0$, then for all $\mathcal{F}_t$-measurable random variables α taking values in $[0, 1]$, we have $\mathcal{E}(\alpha X|\mathcal{F}_t) \geq \mathcal{E}(X|\mathcal{F}_t)$.

7.2.2 Information structures for Multi-armed bandit

We now consider combining play over multiple arms.

In order to retain consistency for a single-arm, the nonlinear expectation needs to be defined together with the filtration. It follows that we need to define an 'independent' nonlinear expectation on the joint space of the arms, which we do via an orthogonal product space. This restriction does not allow us to directly implement Mandelbaum's [77] original approach for a dynamic allocation strategy (Definition 6.5). This is because the multi-parameter process $(\tilde{\eta}(n))$ is only defined to be measurable with respect to the orthant filtration. In particular, it is not clear how one could directly extract the component of $(\tilde{\eta}(n))$ to the marginal space $\Omega^{(k)}$ where our single-bandit nonlinear expectation is defined.

The importance of decomposing a strategy on the multi-armed bandit to strategies for a single-armed bandits can be seen in the proof of El Karoui and Karatzas [63, Equation 5.1] (via Whittle's approach [110]), and is described more explicitly in their continuous time paper [42, Equation 6.9].

In order to overcome this difficulty, we introduce a class of allocation strategies where there is a component associated with the stopping times of the marginal filtrations. This component allows us to connect and separate the space of multiple arms to the marginal space of each arm.

Our class of allocation strategies consists of two components (τ, p). The collection of random times $\tau = (\tau_i^{(k)})_{i\geq 0, k\in\{1,\dots,K\}}$ will identify the duration for which will play the kth arm, the ith time we start to play. This sequence is chosen based on historical observations of the kth arm only, that is, the random times $\sum_{i=0}^{L}\tau_i^{(k)}$ are $(\mathcal{F}_t^{(k)})_{t\geq 0}$-stopping times for all $L \geq 0$. Once we play an arm for $\tau_i^{(k)}$ trials, we will then reconsider which arm to play. Our choice of new arm (which may be the same as before) will be described by the sequence (p_n) taking values in $[K] = \{1, \dots, K\}$, and may depend on information from all arms. The allocation strategy can be defined formally as follows:

Definition 7.2. We say $\tau := \left(\tau_i^{(k)}\right)_{i\geq 0, k\in[K]}$ is a *family of time allocation sequences* if

(i) For each k, $(\tau_i^{(k)})_{i\geq 0}$ is a sequence of non-negative random times defined on the space $(\Omega^{(k)}, \mathcal{F}_\infty^{(k)})$.

(ii) $\sum_{i=0}^{l}\tau_i^{(k)}$ is an $(\mathcal{F}_t^{(k)})_{t\geq 0}$-stopping time for all $l \geq 0$.

Intuitively, the random sequence (p_n) is allowed to depend on all prior observations from all arms. For the sake of precise bookkeeping we need to record, at each moment, how many times we have already played each arm. This leads to the following definition.

Definition 7.3. Given a family of time allocation sequences τ, we say a sequence of random variables $(\eta_n)_{n\geq 0}$ taking values in $\mathcal{S} = \mathbb{N}_0^K$ is a *recording sequence* associated to τ, with corresponding *choice sequence* $(p_n)_{n\in\mathbb{N}_0}$ taking values in $\{0\} \cup [K]$, if

(i) $\eta_0 = (0, \dots, 0)$.

(ii) $\eta_{n+1} = \eta_n + e^{(p_n)}$.

The choice process p_n satisfies

(iii) for all $k \in [K]$ and $r \in \mathcal{S}$,

$$\{p_n = k\} \cap \{\eta_n = r\} \in \mathcal{F}(\Psi_r) = \bigotimes_{k=1}^{K} \mathcal{F}^{(k)}_{\Psi_r^{(k)}}$$

where $\Psi_r^{(k)} := \sum_{i=0}^{r^{(k)}-1}\tau_i^{(k)}$. In particular, $\{\eta_n = r\} \in \mathcal{F}(\Psi_r)$.

For a given time allocation sequence τ, the recording sequence η_n determines the *decision filtration*, given by

$$\mathcal{G}_n^{(\tau,p)} := \left\{A \in \mathcal{F}(T) \,:\, A \cap \{\eta_n = r\} \in \mathcal{F}(\Psi_r) \quad \forall r \in \mathcal{S}\right\} \tag{7.2.6}$$

where $\Psi_r^{(k)} = \sum_{i=0}^{r^{(k)}-1}\tau_i^{(k)}$.

Remark 7.4. We can see in Definition 7.3(iii) that (p_n) is adapted to the filtration $(\mathcal{G}_n^{(\tau,p)})_{n\geq 0}$, i.e. we have made our decision what to do next based on our previous observations.

Definition 7.4. An *(admissible) allocation strategy* (τ, p) consists of a family of time allocation sequences τ and a $(\mathcal{G}_n^{(\tau,p)})_{n\geq 0}$-adapted choice sequence p (defined under τ).

Example 7.1. Suppose there are two arms. The first arm gives only 2 outcomes: $\{w, l\}$. Consider the strategy of playing the first arm until we see the first l. Then we swap to the second arm for two trials and swap back to the first arm and repeat the same procedure.

In this case, we define $(X_t)_{t\geq 1}$ to be the outcome of the first arm and define $\theta_{k+1} := \inf\{t \geq 1 \,:\, X_{t+\sum_{i=0}^{k}\theta_i} = l\}$. We then have the representation of this strategy

$$\tau^{(1)} = (\theta_0, \theta_1, ...), \quad \tau^{(2)} = (2, 2, 2, ...), \quad \text{and} \quad p = (1, 2, 1, 2, ...).$$

The corresponding recording sequence is

$$\eta = \big((0,0), (1,0), (1,1), (2,1), (2,2), (3,2), ...\big).$$

The same strategy can be represented in multiple ways. Here, for example, we can also write

$$\tau^{(1)} = (\theta_0, \theta_1, ...), \quad \tau^{(2)} = (1, 1, 1, ...), \quad \text{and} \quad p = (1, 2, 2, 1, 2, 2, 1, ...).$$

The corresponding recording sequence becomes

$$\eta = \big((0,0), (1,0), (1,1), (1,2), (2,2), (2,3), (2,4), (3,4), (3,5), (3,6), (4,7), ...\big).$$

As discussed in Remark 6.7, we can express a Mandelbaum allocation strategy (Definition 6.5) in terms of a sequence ρ of decisions made at each time. For the strategy described above, this gives the unique sequence

$$\rho = \big(\underbrace{1, 1, .., 1}_{\theta_0}, 2, 2, \underbrace{1, 1, .., 1}_{\theta_1}, 2, 2, \underbrace{1, 1, .., 1}_{\theta_2}, ...\big).$$

Extending this example, we can generally write our strategy (τ, p) in terms of ρ and vice versa. This unique representation provides a simple (if inefficient) description of our strategy, which we now make precise.

Definition 7.5. Define the random variable ρ_n to be the arm which will be observed in the nth play under an admissible allocation strategy (τ, p). We call the process $(\rho_n)_{n\geq 0}$, a *simple form* allocation sequence. The construction of the sequence (ρ_n) is given explicitly in Lemma B.5 in the appendix.

For admissible allocation strategies (τ, p) and $(\hat{\tau}, \hat{p})$, we write $(\tau, p) \sim (\hat{\tau}, \hat{p})$ if they lead to the same simple form. (Clearly, $\sim$ defines equivalence classes.)

Remark 7.5. Observe that if ρ is the simple form of (τ, p) and we denote the time allocation sequence

$$\mathbf{1} = \big(1, 1, 1, ...\big), \tag{7.2.7}$$

then $(\mathbf{1}, \rho)$ is an allocation strategy which yields the same decisions as (τ, p). In particular, we have $(\mathbf{1}, \rho) \sim (\tau, p)$.

Furthermore, one can check that the recording sequence corresponding to $(\mathtt{1}, \rho)$ is exactly the Mandelbaum allocation strategy (Definition 6.5). In particular, we can explicitly construct a one-to-one correspondence between our equivalence class of admissible strategies (Definition 7.4) and Mandelbaum allocation strategies, and we have $\mathcal{G}_n^{(\mathtt{1},\rho)} = \mathcal{G}_n^{\rho}$ in (6.4.3).

Remark 7.6. Assume that, for $k \in [K]$, the filtration $(\mathcal{F}_t^{(k)})$ is generated by an underlying real process $(\xi_t^{(k)})_{t\geq 1}$ defined on the space $(\Omega^{(k)}, \mathcal{F}_\infty^{(k)})$. i.e.

$$\mathcal{F}_0^{(k)} = \left\{\phi, \Omega^{(k)}\right\} \quad \text{and} \quad \mathcal{F}_t^{(k)} = \sigma\big(\xi_1^{(k)}, \xi_2^{(k)}, ..., \xi_t^{(k)}\big).$$

If we parameterise our actions by a simple form strategy $(\mathtt{1}, \rho)$ with associated recording sequence η, then ρ_{n-1} is the decision made at time $n-1$ to generate the outcome observed at time n. The observation at the nth play is given by

$$\xi_n^{\rho} := \xi_{\eta_n^{(\rho_{n-1})}}^{(\rho_{n-1})} = \sum_{m=1}^{M}\sum_{t=1}^{\infty} \xi_t^{(k)} \mathbb{I}(\rho_{n-1} = k,\ \eta_n^{(k)} = t).$$

We define the *observed filtration* by $\mathcal{H}_0^{\rho} = \{\phi, \bar{\Omega}\}$ and

$$\mathcal{H}_n^{\rho} := \sigma\left(\xi_1^{\rho}, ..., \xi_n^{\rho}\right) \quad \text{for} \quad n = 1, 2, ... \tag{7.2.8}$$

We prove, in the appendix, that the observed filtration agrees with that used in Definition 7.3 when considering measurability of ρ. That is

$$\mathcal{H}_n^{\rho} = \mathcal{G}_n^{(\mathtt{1},\rho)} = \{A \in \mathcal{F}(T) \ : \ A \cap \{\eta_n = r\} \in \mathcal{F}(r)\ \} \tag{7.2.9}$$

where η is the recording sequence corresponding to $(\mathtt{1}, \rho)$.

Multi-armed bandit optimality

We can now give the second half of the proof for the robust Gittins index theorem.

In order to prove the optimality of the robust Gittins' strategy, we define the target function for an allocation strategy by

$$V(\tau, p) := \mathfrak{E}\left(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1})}(t_n^{\rho}) - \Gamma^{(\rho_{n-1})}(t_n^{\rho}) \right) \right) \ : \ t_n^{\rho} := \sum_{i=0}^{n-1} \mathbb{I}(\rho_i = \rho_{n-1}) \tag{7.2.10}$$

where ρ is a simple form derived from (τ, p) and $\left(\Gamma^{(k)}(t)\right)$ is the running max of the robust Gittins index of the kth arm, as considered in (7.2.3).

Step B.1 Suppose that we have K arms, with associated indifference rewards $(\Gamma^{(k)})_{k\in[K]}$ as in step A.3. If we mix the play of these arms, this is equivalent to taking a break in a single arm to play the others. This delay will result in a possibly worse outcome (Eq. (7.2.5) in step A.4).

In Theorem 8.12, we use the definition of $\mathfrak{E}$ and apply Fubini's theorem to show that, for all allocation strategies (τ, p), this implies that for any $\epsilon > 0$, there exists a probability measure $\bigotimes_{m=1}^{M} \mathbb{Q}^{(k)} \in \mathcal{Q}$ such that

$$\begin{aligned} V(\tau,p) &\geq \mathbb{E}^{\bigotimes_{m=1}^{M}\mathbb{Q}^{(k)}}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho})-\Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right) \\ &= \sum_{m=1}^{M}\mathbb{E}^{\mathbb{Q}^{(k)}}\left(\sum_{t=1}^{\infty}\tilde{\alpha}^{(k)}(t)\beta^t\left(h^{(k)}(t)-\Gamma^{(k)}(t)\right)\right) \quad \geq \quad -M\epsilon \end{aligned}$$

where $\tilde{\alpha}^{(k)}(t)$ is the delay effect on the mth arm due to playing other arms.

As ϵ is arbitrary, it follows that for all allocation strategies (τ, p),

$$V(\tau,p) \geq 0. \tag{7.2.11}$$

Step B.2 In step A.2, we noticed that the total expected loss of a single arm between S and S' is zero, for S and S' the consecutive stopping times when the robust Gittins index hits a new maximum (7.2.2). We use this fact to construct a family of time allocation sequences as a candidate for an optimal strategy.

Define (inductively) $S_i^{(k)} := \sum_{l=0}^{i-1}\sigma_l^{(k)}$ and

$$\sigma_i^{(k)} := \inf\left\{\theta \geq 1 : \gamma^{(k)}\left(S_i^{(k)}+\theta\right) > \gamma^{(k)}\left(S_i^{(k)}\right)\right\} \tag{7.2.12}$$

Using our construction on the class of allocation strategies, we can project the joint nonlinear valuation to its marginal space which is equipped with a consistent nonlinear expectation. We can then use the subadditivity of $\mathfrak{E}$ and the result from step A.2, that $\sigma_i^{(k)}$ yields equality in (7.2.1), to show that, for any p, with the choice of time allocation sequences $\sigma = (\sigma_i^{(k)})$, the allocation strategy (σ, p) has value

$$V(\sigma,p) \leq 0.$$

This result is expressed and proved in Theorem 8.13.

Step B.3 By combining step B.1 and step B.2, for any p, with the choice of time allocation sequences $\sigma = (\sigma_i^{(k)})$ considered above, we have

$$V(\sigma,p) = 0. \tag{7.2.13}$$

We consider

$$C^{\rho}(n) = \begin{cases} \beta\Gamma^{(\rho_0)}(t_1^{\rho}) + \mathfrak{E}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho})-\Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right) & \text{for } n=1 \\ \beta^n\Gamma^{(\rho_{n-1})}(t_n^{\rho}) & \text{for } n \geq 2, \end{cases}$$

and $Y^{\rho} := \sum_{n=1}^{\infty} C^{\rho}(n)$ as compensators in Definition **??** and 6.9.

The strategy ρ^* given in Theorem 7.2 is the strategy of always playing the arm with the minimal index. Therefore, it lies in the same equivalence class as a strategy

with the time allocation sequences $(\sigma^{(k)})$ (and with p indicating the minimum index amongst all arms at each time). Hence, by (7.2.13),

$$\mathfrak{E}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho^*_{n-1})}(t^{\rho}_n)-\Gamma^{(\rho^*_{n-1})}(t^{\rho^*}_n)\right)\right)=0.$$

By monotonicity of the process Γ, we prefer lower value earlier, due to the discount effect. By using this observation together with (7.2.11), we can show that the strategy ρ^* satisfies the optimality conditions (**??**) and (**??**) in Definition **??** and 6.9.

Finally, as Γ is a running maximum, restarting our analysis could only yield a reduction in Γ. From this observation, we can then establish condition (**??**) in Definition 6.9.

Hence, ρ^* must be (static/dynamic) C-optimal. The formal proof is given in Theorem 8.14.

7.3 Numerical evaluation of Robust Gittins indices

In this section, we study the behaviour of the robust Gittins index through a numerical example. Again we omit the superscript (k) for notational simplicity.

We suppose the arm under consideration generates independent identically distributed costs $(h(t))_{1\le t\le T}$ of either \$1 or \$0, given (unknown) probability $\mathbb{P}(h(t)=1)=\theta$ and $h(t)=2$ for all $t>T$. The filtration $(\mathcal{F}_t)_{t\ge 0}$ is generated by the observed cost process $(\xi_t)_{1\le t\le T}=(h(t))_{1\le t\le T}$ (with $\mathcal{F}_0$ trivial). The horizon T can be thought of as the maximum number of times that each arm can be played.

Remark 7.7. An imaginary horizon T is introduced to allow us to easily construct a data-driven recursive nonlinear expectation (7.3.1) by backward induction.

The future cost $h(t)=2$ is introduced to simplify our numerical method. By considering (7.1.1), we can see that the robust Gittins index $(\gamma(t))_{t\ge 1}$ takes values between 0 and 1 when $t\le T$ and $\gamma(t)=2$ for $t>T$. Moreover, the optimal stopping time $\sigma(t,\gamma(t))\le T-t$. Hence, one can calculate the robust index $\gamma(t)$ by considering a finite horizon optimal stopping problem.

We model uncertainty in this setting by constructing a one-step coherent nonlinear expectation (inspired by the DR-Expectation [29], see also Bielecki, Chen and Cialenco [16]) given by

$$\mathcal{E}_{(t)}\Big(f(\xi_1,...,\xi_t,\xi_{t+1})\Big):=\sup_{\theta\in\Theta_t}\Big(\theta f(\xi_1,...,\xi_t,1)+(1-\theta)f(\xi_1,...,\xi_t,0)\Big)\tag{7.3.1}$$

where $\Theta_t=\Big[p^-(p_t,n_t),p^+(p_t,n_t)\Big]$ corresponds to a credible interval for θ given our observations at time t, using a (nearly improper) Beta prior distribution. The processes n_t and p_t correspond to the number of observations and the (posterior mean) estimate of θ at time t.

In particular, we may choose a credible level $\alpha\in[0,1]$ and obtain $p^{\pm}(p_t,n_t)$ by

$$p^{\pm}(p_t,n_t)=I^{-1}_{(p_tn_t+\epsilon,(1-p_t)n_t+\epsilon)}(0.5\pm\alpha/2)$$

where $q \mapsto I^{-1}_{(a,b)}(q)$ is the quantile function of the Beta(a,b) distribution and $\epsilon = 10^{-6}$ is a small quantity added to avoid degeneracy.

Remark 7.8. One could also use the central limit theorem to obtain an asymptotic confidence interval. However, because $\Theta_t \subseteq [0,1]$, we restrict ourselves to the credible set above to avoid end-effects, and allow for asymmetry in the plausible values around the 'best' estimate.

By Proposition 6.1, $\mathcal{E}_{(t)}$ induces an $(\mathcal{F}_t)_{t\geq 0}$-consistent coherent nonlinear expectation $\mathcal{E}$ by backward recursion defined up to a finite horizon T. As our credible set is constructed from p_t and n_t, and the pair (p_t, n_t) can be computed recursively, it follows that for every $f : \{0,1\}^{T-t} \to \mathbb{R}$, there exists a function $g_{\alpha,T-t} : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathcal{E}\big(f(\xi_{t+1}, ..., \xi_T)\big|\mathcal{F}_t\big) = g_{\alpha,T-t}\Big(p_t, \frac{1}{\sqrt{n_t}}\Big).$$

Remark 7.9. Here, we write the nonlinear expectation as a function of $n_t^{-1/2}$ instead of n_t as we wish to approximate our function on a compact domain. The choice of $n_t^{-1/2}$ comes from the natural scaling of the credible set.

Now, recall that

$$\gamma(s) := \inf\left\{\gamma \in L^\infty(\mathcal{F}_s) \ : \ \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} \mathcal{E}\left(\sum_{t=1}^{\tau} \beta^t \left(h(s+t) - \gamma\right) \ \middle| \ \mathcal{F}_s\right) \leq 0\right\}$$

where $\mathcal{T}(s)$ is the space of positive $(\mathcal{F}_{s+t})_{t\geq 0}$-stopping times.

By following a general robust dynamic programming argument, as in Ruszczyński [99], or using the nonlinear Snell's envelope, as in Riedel [90], it follows that we can write

$$\gamma(t) = \gamma_{\alpha,\beta,T-t}\Big(p_t, \frac{1}{\sqrt{n_t}}\Big) \qquad \text{where} \qquad n_t = n_0 + t.$$

for some function $\gamma_{\alpha,\beta,T-t}$.

We then use a simple finite-difference algorithm (see Appendix B.3) to estimate the function

$$\Big(p, \frac{1}{\sqrt{n}}\Big) \mapsto \gamma_{\alpha,\beta,T-(n-n_0)}\Big(p, \frac{1}{\sqrt{n}}\Big) - p$$

where, in our simulations, we fix $n_0 = 1$.

Plots of this estimate, for various values of α, β and T, can be found in Figure 7.1 where the case $T = 10$ is truncated as n cannot exceed T (by definition).

For $h(t) = \xi_t$, with uncertainty modeled by (7.3.1), at each time step we wish to play the arm with the lowest θ. Classically, this is estimated by p, so a naïve (greedy) strategy would suggest playing the arm with the lowest estimated average loss p. By using C-optimality, at each point, we choose an arm with the lowest γ. Therefore, we may think of γ as an implied probability p, distorted to account for exploration and exploitation of the system of arms, i.e., a robust analogy of incremental value (3.0.1) considered in Chapter 3.

In Figure 7.1, we see the following broad phenomena:

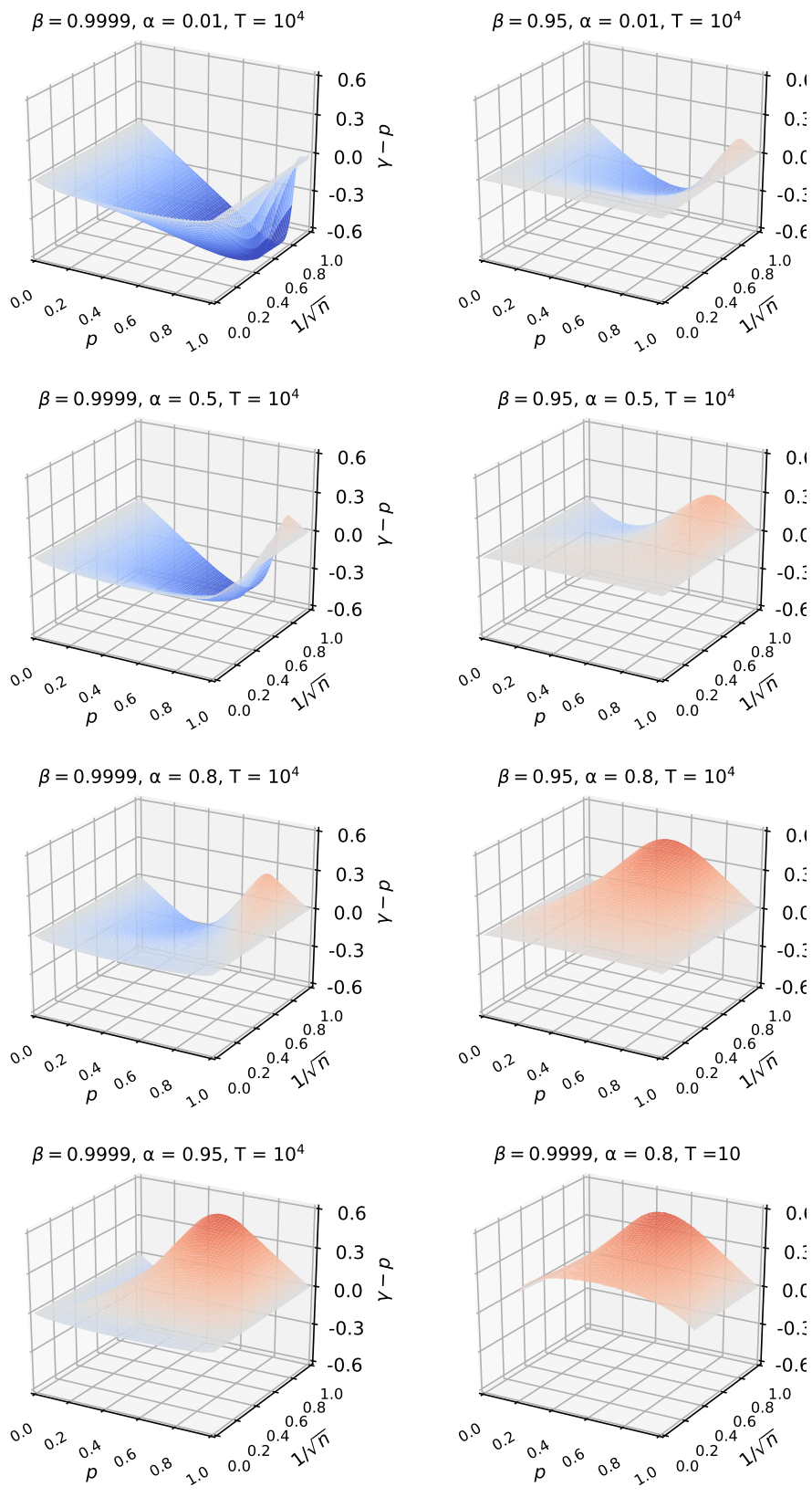

Figure 7.1: Estimated value of $\gamma - p$ for different values of α, β and T

- When $1/\sqrt{n}$ is small, the difference between γ and p is close to zero. In particular, this says that when we have high certainty in our estimates, γ is equivalent to the estimated probability.

- when β is low, $\gamma - p$ typically becomes positive when n is large. This means that as n gets large, we tend to be more uncertainty aversive than be active to learn. We will explain this observation more rigorously in Section 9.3.

- When we increase β, the difference $\gamma - p$ typically decreases. This corresponds to the fact that β is a discount factor, which determines how much we value future costs. Therefore, increasing β increases the degree that we wish to explore the system, i.e. we become more optimistic in our evaluation. We also observe that decreasing β also yields a similar result to shortening the horizon.

- When α increases, the difference $\gamma - p$ increases. This is due to the fact that α corresponds to the 'width' of the 'credible interval'. Hence, large α means that we become more conservative and favour exploiting over exploring.

7.3.1 Prospect Theory

One result suggested in Figure 7.1 when $\beta = 0.9999$ and $\alpha = 0.01$ is that, when we do not worry about uncertainty, we are more optimistic when p is large (close to 1), that is, γ is clearly less than p. On the other hand, when uncertainty dominates, e.g. when $\beta = 0.9999$ and $\alpha = 0.95$, or $\beta = 0.95$ and $\alpha = 0.8$, we become more pessimistic.

Curiously, when $\beta = 0.9999$ and $\alpha = 0.8$, or $\beta = 0.95$ and $\alpha = 0.5$, both optimism and pessimism can be seen. For large p, (when the game seems bad), pessimism dominates, while for small p (when the game seems good), we become optimistic in our optimal strategy. This gives a bias in the probabilities, related to that used in the probability weighting functions as considered in prospect theory by Kahneman and Tversky [61] or in rank-dependent expected utility by Quiggin [86, 87]. In this literature, they propose models to explain irrationality in human decisions under risk. They argue that people generally reweigh the probabilities of different outcomes using a nonlinear increasing map $p \mapsto \pi(p)$, with various assumptions on its curvature.

Our result (for appropriate values of β and α) reflects this behaviour without imposing a probability weighting function as in classical prospect theory. Instead, the combination of the effect of learning and uncertainty leads to distortions of the estimated probability.

7.4 Monte-Carlo simulation

In order to illustrate the effect of the uncertainty aversion to the performance of robust Gittins index, we consider the Bernoulli bandit as described above over 50 exchangeable arms and for a horizon $T = 10^4$. We run 10^3 Monte-Carlo simulations and compare the performance of various strategies for decision making.

To provide a wide range of scenarios in which our strategies must perform, in each simulation, we first generate a, b independently from a $\Gamma(1, 1/100)$ distribution, then generate the 'true' probabilities for each arm independently from $\text{Beta}(a, b)$. We generate 10 trials on each arm to provide initial information.

In these simulations, we will focus on the effect of the discount factor and uncertainty aversion to the robust Gittins index. Therefore, we will consider UCB (with $\lambda = 1$), Thompson sampling (with prior $\text{Beta}(1, 1)$), and greedy algorithm as a benchmark for the simulation. We will call the algorithm corresponding to the robust Gittins index the DR (Data-Robust) algorithm.

Remark 7.10. To avoid bias in the algorithms, we choose an arm uniformly at random if there is more than one arm with minimal index.

7.4.1 Measures of Regret

To see the effect of the discount factor and uncertainty aversion on the performance, we will consider the (pseudo) expected–expected regret as in (4.3.1). To explain the decision that the robust Gittins index made, we will also consider the number of sub-optimal plays.

- **Expected–expected regret:** This is the difference in the true expectations under our strategy and an optimal strategy with perfect information. In our setting, this can be given by

$$R(\rho, T, \Theta) = \sum_{n=0}^{T} (\Theta^{(\rho_n)} - \Theta^*).$$

  where $\Theta^{(k)}$ is the true probability of the kth arm and $\Theta^* = \min_k \Theta^{(k)}$.

- **Sub-optimal plays.** This measures the number of times where we play a sub-optimal arm.

$$N_v(\rho, T, \Theta) = \sum_{n=0}^{T} \mathbb{I}(\Theta^{(\rho_n)} \neq \Theta^*).$$

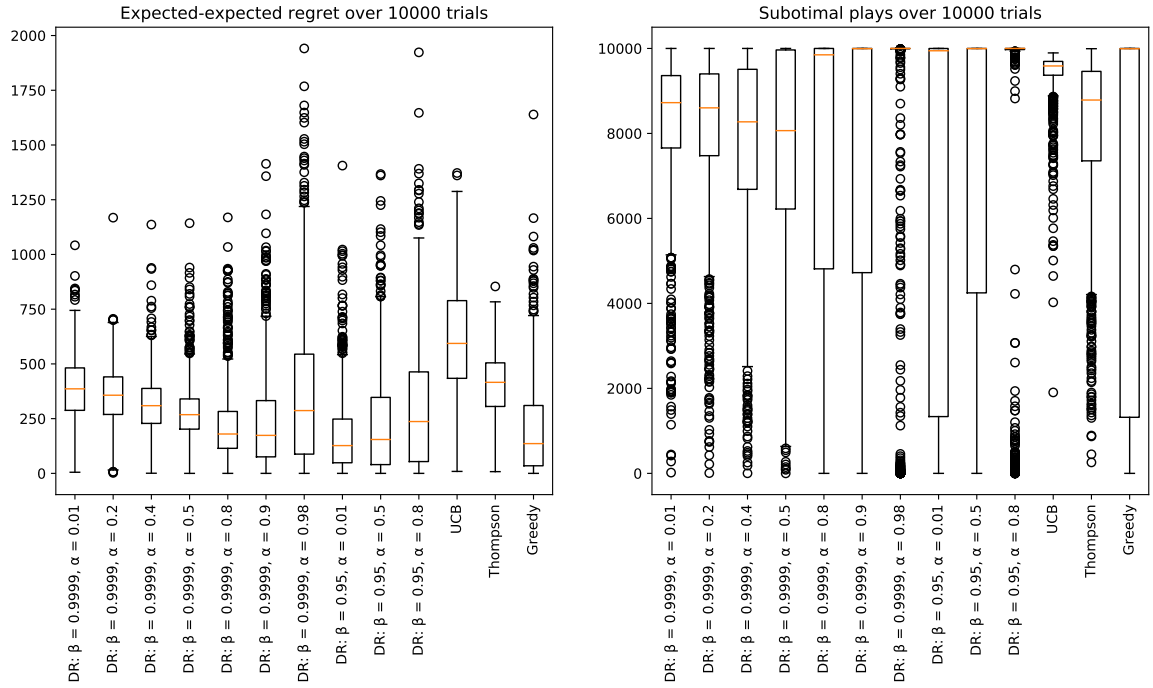

Figure 7.2: Regret for Bernoulli bandit under different policies

In Figure 7.2, considering first the cases where $\beta = 0.9999$, we can see that an increase in the value of α has a nonlinear effect on the distribution of regret. Initially, increasing α appears to reduce the typical regret and suboptimal plays. However, setting α too large clearly leads to worse outcomes. This is because α corresponds to the level of robustness; the more robust we are, the less willing we are to explore and the more willing we are to exploit. It follows that a large value of α encourages us to exploit early, and we may not find the optimal arm to play.

On the other hand, the discount rate β determines how much we value our future costs. If we have a high level of robustness (large α) but do not value the future cost enough (small β), we may end up settling for a sub-optimal decision. This can be seen most clearly when $\beta = 0.9999$ and $\alpha = 0.8$. In this case, the average expected-expected regret is relatively small compared to other strategies, but its average number of suboptimal plays is relatively high. Reducing β to 0.95 emphasises these effects even further.

7.4.2 Robustness of the DR Algorithms

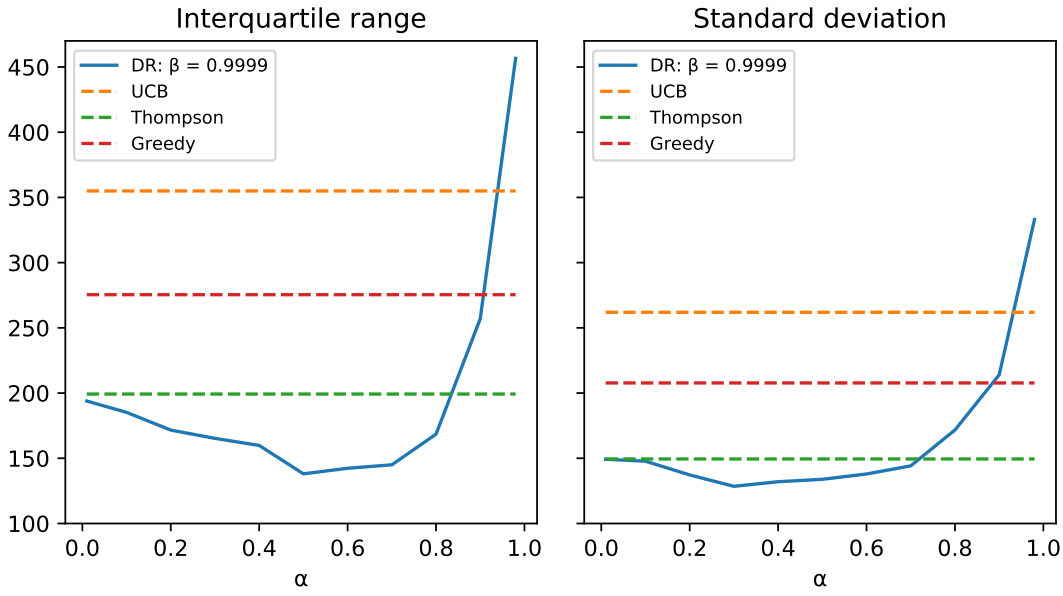

Figure 7.3: Deviation of the expected-expected regret when $\beta = 0.9999$

In Figure 7.3, we illustrate the interquartile range and the standard deviation of the total expected-expected regret when $\beta = 0.9999$ with different values of α over 1000 simulations. We can see that by introducing an appropriate values of α, we can obtain a substantial reduction in the interquartile range and the standard deviation. In particular, the DR algorithm with an appropriate uncertainty aversion level (α) does not only give a low average regret but also does so consistently over different simulations.

Nonetheless, introducing too much uncertainty aversion level (α) without having a sufficient willingness to explore (β) could result in a wide range of regret, as we observed when α is large.

Chapter 8

Proof of Gittins Theorem under uncertainty aversion

In this chapter, we will flesh out the sketch given in Chapter 7. The proof is split into two parts. The first half of the proof contains a careful analysis of an optimal stopping problem under nonlinear expectation, and the corresponding 'fair value' process, for a single-armed bandit. The second half of the proof combines the arms together, and demonstrates that the single-arm analysis yields an optimal strategy when deciding between multiple arms. Further technical lemmas, which are used but do not contribute significantly to the main proof, are given in Appendix B.

This chapter is based on the paper [35].

8.1 Part A: Analysis of a single arm

In this section, we will focus the discussion on a single-armed bandit.

8.1.1 Step A.1: Indifference reward and Optimal Stopping problem

We first recall the definition of the robust Gittins index (process).

$$\gamma(s) := \operatorname{ess\,inf}\left\{\gamma \in L^\infty(\mathcal{F}_s) : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} \mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t\big(h(t)-\gamma\big) \,\middle|\, \mathcal{F}_s\right) \le 0\right\}$$

where $\mathcal{T}(s)$ denotes the family of $(\mathcal{F}_{s+t})_{t\ge 0}$-positive stopping times.

Remark 8.1. If we take $\tau = 1$, we observe by boundedness of h (Assumption 6.2) that $\gamma(t) < C$ where C is an upper bound of h.

To study the process γ, we introduce an auxiliary optimal stopping problem. At each time step, the player decides whether to continue or to stop play of the machine. If the player decides to continue to play, he will be offered a fixed reward λ (known at the initial time s) in addition to the cost $h(t)$.

Definition 8.1. The target function $V_s : \mathcal{T}(s) \times L^\infty(\mathcal{F}_s) \to L^\infty(\mathcal{F}_s)$ for a stopping time $\tau \in \mathcal{T}(s)$ with a reward λ is defined by

$$V_s(\tau, \lambda) = \mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \;\middle|\; \mathcal{F}_s \right).$$

We know that $\gamma(s)$ is defined to be the minimum reward λ such that, with a choice of τ minimizing $V_s(\tau, \lambda)$, the expected loss is at most zero. By minimality of $\gamma(s)$ and monotonicity of $\mathcal{E}$, the reward $\gamma(s)$ will yield zero loss under optimal stopping and, therefore, cannot yield a positive expected reward under suboptimal stopping. In particular, the following holds.

Theorem 8.1. *The function V_s defined above satisfies.*

$$\operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \gamma(s)) = 0.$$

Corollary 8.2. *For every $\tau \in \mathcal{T}(s)$, we have*

$$\mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \gamma(s)\big) \middle| \mathcal{F}_s \right) \geq 0.$$

Remark 8.2. Theorem 8.1 shows that, under optimal stopping, with the reward $\gamma(s)$, the expected total loss is zero. In particular, we may view $\gamma(s)$ as an 'average cost under optimal play' of the arm.

8.1.2 Step A.2: Optimal Stopping time

By considering a Snell envelope argument, as in Riedel [90] with slight modification, we can establish that there exists a stopping time τ^* achieving the minimum value $V_s(\tau^*, \lambda) = \operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \lambda)$ exists (Theorem B.4). In this subsection, we will show that τ^* can be expressed as a hitting time of the Gittins index process $(\gamma(s))$.

Definition 8.2. Let λ be a non-negative $\mathcal{F}_s$-measurable random variable. Define a stopping time $\sigma(s, \lambda)$ by

$$\sigma(s, \lambda) := \inf\{\theta \geq 1 \,:\, \gamma(s + \theta) > \lambda\}$$

As mentioned in Remark 8.2, we may view γ as a time-average cost under optimal stopping. The stopping time $\sigma(s, \lambda)$ can be interpreted as the first time when this average cost exceeds a fixed λ. Once γ exceeds λ, the offered compensation λ is insufficient to make the arm attractive; so, to minimise the total 'expected' cost, we will stop.

In what follows, we formalise this intuition. We will show that $\sigma(s, \lambda)$ is an optimal stopping time when the reward λ is offered. In particular, we will show that $\sigma(s, \gamma(s))$ attains the optimal value with the reward $\lambda = \gamma(s)$. Moreover, the value for this optimal stopping problem is zero (by Theorem 8.1).

The optimality of $\sigma(s, \lambda)$ can be proved by showing that for any stopping time $\tau \in \mathcal{T}(s)$, if $\tau > \sigma(s, \lambda)$ on some event, our value can be improved by stopping at $\sigma(s, \lambda)$ (Lemma 8.3). On the other hand, if $\tau < \sigma(s, \lambda)$ on some event, the value can be improved by continuing to play (Lemma 8.4).

Lemma 8.3. *For every $\lambda \in L^\infty(\mathcal{F}_s)$ taking values in $[0, C)$ and $\tau \in \mathcal{T}(s)$,*

$$V_s(\tau, \lambda) \geq V_s(\tau \wedge \sigma(s, \lambda), \lambda).$$

Proof. We will prove this result by applying Corollary 8.2 together with time-consistency and monotonicity of our nonlinear expectation.

Define $\nu = \tau \wedge \sigma(s, \lambda)$. By Corollary 8.2 and regularity (Remark 6.3),

$$\begin{aligned}
0 &\leq \mathcal{E}\left(\sum_{t=s+\nu+1}^{s+\tau} \beta^t \left(h(t) - \gamma(s+\nu)\right) \middle| \mathcal{F}_{s+\nu} \right) \\
&= \mathbb{I}(\tau > \sigma(s,\lambda)) \mathcal{E}\left(\sum_{t=s+\nu+1}^{s+\tau} \beta^t \left(h(t) - \gamma(s+\sigma(s,\lambda))\right) \middle| \mathcal{F}_{s+\nu} \right) + \mathbb{I}(\tau \leq \sigma(s,\lambda))(0) \\
&\leq \mathbb{I}(\tau > \sigma(s,\lambda)) \mathcal{E}\left(\sum_{t=s+\nu+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\nu} \right) = \mathcal{E}\left(\sum_{t=s+\nu+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\nu} \right).
\end{aligned}$$

By translation equivariance,

$$\begin{aligned}
\sum_{t=s+1}^{s+\nu} \beta^t \left(h(t) - \lambda\right) &\leq \sum_{t=s+1}^{s+\nu} \beta^t \left(h(t) - \lambda\right) + \mathcal{E}\left(\sum_{t=s+\nu+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\nu} \right) \\
&= \mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\nu} \right).
\end{aligned}$$

By monotonicity and time-consistency,

$$\begin{aligned}
\mathcal{E}\left(\sum_{t=s+1}^{s+\nu} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_s \right) &\leq \mathcal{E}\left(\mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\nu} \right) \middle| \mathcal{F}_s \right) \\
&= \mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_s \right).
\end{aligned}$$

In particular, $V_s(\tau \wedge \sigma(s, \lambda), \lambda) = V_s(\nu, \lambda) \leq V_s(\tau, \lambda)$. □

Lemma 8.4. *Let $\tau \in \mathcal{T}(s)$ and let $\lambda \in L^\infty(\mathcal{F}_s)$ taking values in $[0, C)$ where C is an upper bound of h in Assumption 6.2. Then there exists a stopping time $\tau_1 \in \mathcal{T}(s)$ with $\tau_1 \geq \tau$ such that*

$$V_s(\tau, \lambda) \geq V_s(\tau_1, \lambda)$$

and on the event $A := \{\gamma(s+\tau) \leq \lambda\}$, we have $\tau_1 > \tau$.

Proof. For $A = \{\gamma(s+\tau) \leq \lambda\}$, define $\gamma^\tau := \gamma(s+\tau)\mathbb{I}_{A^c} + \lambda \mathbb{I}_A \geq \gamma(s+\tau)$. By Theorem 8.1 and monotonicity of our nonlinear expectation,

$$0 = \operatorname*{ess\,inf}_{\tilde{\tau} \in \mathcal{T}(s+\tau)} \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\tau+\tilde{\tau}} \beta^t \left(h(t) - \gamma(s+\tau)\right) \middle| \mathcal{F}_{s+\tau} \right) \geq \operatorname*{ess\,inf}_{\tilde{\tau} \in \mathcal{T}(s+\tau)} \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\tau+\tilde{\tau}} \beta^t \left(h(t) - \gamma^\tau\right) \middle| \mathcal{F}_{s+\tau} \right).$$

Thus, by Theorem B.4, there exists $\tilde{\tau}^* \in \mathcal{T}(s+\tau)$ such that,

$$0 \geq \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\tau+\tilde{\tau}^*} \beta^t \left(h(t) - \gamma^\tau\right) \middle| \mathcal{F}_{s+\tau} \right).$$

Define a stopping time $\tau_1 := \tau + \tilde{\tau}^* \, \mathbb{I}_A$. As $\gamma^\tau \mathbb{I}_A = \lambda \mathbb{I}_A$, then

$$\begin{aligned}\sum_{t=s+1}^{s+\tau_1} \beta^t \left(h(t) - \lambda\right) &= \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) + \mathbb{I}_A \sum_{t=s+\tau+1}^{s+\tau+\tilde{\tau}^*} \beta^t \left(h(t) - \lambda\right) \\ &= \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) + \mathbb{I}_A \sum_{t=s+\tau+1}^{s+\tau+\tilde{\tau}^*} \beta^t \left(h(t) - \gamma^\tau\right).\end{aligned}$$

By translation equivariance and regularity (Remark 6.3), it follows that

$$\begin{aligned}\mathcal{E}\left(\sum_{t=s+1}^{s+\tau_1} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right) &= \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) + \mathbb{I}_A \mathcal{E}\left(\sum_{t=s+1}^{s+\tau+\tilde{\tau}^*} \beta^t \left(h(t) - \gamma^\tau\right) \middle| \mathcal{F}_{s+\tau} \right) \\ &\geq \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right).\end{aligned}$$

Finally, by applying monotonicity and time-consistency as in the previous lemma, the result follows. □

Corollary 8.5. *Let $\tau \in \mathcal{T}(s)$. Then there exists an increasing sequence $(\tau_n)_{n\geq 1}$ in $\mathcal{T}(s)$ with $\tau_{n+1} \geq \tau_n \geq \tau$ for all $n \geq 1$ such that*

$$V_s(\tau, \lambda) \geq V_s(\tau_1, \lambda) \geq \ldots \geq V_s(\tau_n, \lambda)$$

and on the event $\bigcap_{k=1}^{n-1} \{\gamma(s + \tau_k) \leq \lambda\}$ we have $\tau_n > \tau_{n-1} > \ldots > \tau_1 > \tau$. In particular, on this event, $\tau_n \geq n$.

By combining these observations with the Lebesgue property of $\mathcal{E}$, we have the following theorem.

Theorem 8.6. *For every $\lambda \in L^\infty(\mathcal{F}_s)$ taking values in $[0, C)$ and $\tau \in \mathcal{T}(s)$, we have*

$$V_s(\tau, \lambda) \;\geq\; V_s(\sigma(s, \lambda), \lambda) = \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \lambda).$$

Therefore,

$$V_s(\sigma(s, \gamma(s)), \gamma(s)) = \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \gamma(s)) = 0.$$

In particular, $\sigma(s, \gamma(s))$ yields equality in Corollary 8.2.

Proof. By Lemma 8.3 and Corollary 8.5,

$$V_s(\tau, \lambda) \geq V_s(\tau_n, \lambda) \geq V_s(\tau_n \wedge \sigma(s, \lambda), \lambda).$$

Observe that by Corollary 8.5,

$$\{\tau_n < \sigma(s, \lambda)\} = \{\gamma(s + \theta) \leq \lambda \;\; \forall \theta \leq \tau_n\} \subseteq \bigcap_{k=1}^{n-1} \{\gamma(s + \tau_k) \leq \lambda\} \subseteq \{\tau_n \geq n\}.$$

Hence, it follows that $\tau_n \wedge \sigma(s, \lambda) \to \sigma(s, \lambda)$ as $n \to \infty$ and thus

$$\sum_{t=s+1}^{s+\tau_n \wedge \sigma(s,\lambda)} \beta^t \left(h(t) - \lambda\right) \longrightarrow \sum_{t=s+1}^{s+\sigma(s,\lambda)} \beta^t \left(h(t) - \lambda\right) \quad \text{as } n \to \infty \text{ for all } \omega \in \Omega.$$

As h is bounded, it follows from the Lebesgue property of our nonlinear expectation that

$$\mathcal{E}\left(\sum_{t=s+1}^{s+\tau_n \wedge \sigma(s,\lambda)} \beta^t \big(h(t) - \lambda\big) \,\middle|\, \mathcal{F}_s \right) \longrightarrow \mathcal{E}\left(\sum_{t=s+1}^{s+\sigma(s,\lambda)} \beta^t \big(h(t) - \lambda\big) \,\middle|\, \mathcal{F}_s \right).$$

In particular, $V_s(\tau, \lambda) \geq V_s(\sigma(s,\lambda), \lambda)$. □

8.1.3 Step A.3: Fair Game and Prevailing process

Previously, we considered an optimal stopping problem when the Gittins index is offered as compensation for continued play. In this subsection, we consider a 'fair game' when we offer a compensation which is (just) sufficient to encourage us to continue playing the arm. In particular, the compensation increases at each optimal stopping time in order to encourage the agent to continue.

We will first define a sequence of optimal stopping times that we have to consider in order to analyse our (minimal) compensation process.

Definition 8.3. We define $\hat{S}_n$ to be the stopping time where the Gittins index process $(\gamma(s))_{s\geq 0}$ exceeds its running maximum for the nth time. We write σ_n for the duration between $\hat{S}_n$ and $\hat{S}_{n+1}$, that is, σ_n is a random time identifying how long after time $\hat{S}_n$ the process $(\gamma(s))_{s\geq 0}$ hits a new maximum.

More precisely, we define $\hat{S}_n$ and σ_n inductively:

(i) Let $\hat{S}_0 \coloneqq 0$.

(ii) Given $\hat{S}_n$, define

$$\sigma_n \coloneqq \inf\{\theta \geq 1 \,:\, \gamma(\hat{S}_n + \theta) > \gamma(\hat{S}_n)\}$$

and $\hat{S}_{n+1} \coloneqq \hat{S}_n + \sigma_n$.

Equivalently, we can define $\sigma_n \coloneqq \sigma(\hat{S}_n, \gamma(\hat{S}_n))$ as in Definition 8.2.

Definition 8.4. We define the *prevailing reward* process Γ by the running maximum of γ, that is,

$$\Gamma(t) \coloneqq \max_{0\leq\theta\leq t-1} \gamma(\theta).$$

We can then show that the process Γ serves as an indifference reward (process) for our agent, when evaluated from the perspective of one of the stopping times $\hat{S}_n$.

Proposition 8.7. *For all $n \in \mathbb{N}$,*

$$\mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\infty} \beta^t \left(h(t) - \Gamma(t)\right) \middle| \mathcal{F}_{\hat{S}_n} \right) = 0.$$

In particular,

$$\mathcal{E}\left(\sum_{t=1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \right) = 0.$$

Proof. By Theorem 8.6, we have, for all $k \in \mathbb{N}$,

$$0 = \mathcal{E}\left(\sum_{t=\hat{S}_k+1}^{\hat{S}_{k+1}} \beta^t \big(h(t) - \gamma(\hat{S}_n)\big) \middle| \mathcal{F}_{\hat{S}_k} \right).$$

Fix $n, N \in \mathbb{N}$ with $N \geq n$, by time-consistency, translation equivariance,

$$\begin{aligned}
&\mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\hat{S}_N} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\hat{S}_n} \right) \\
&= \mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}} \beta^t \big(h(t) - \Gamma(t)\big) + \mathcal{E}\left(\sum_{t=\hat{S}_{N-1}+1}^{\hat{S}_N} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\hat{S}_{N-1}} \right) \middle| \mathcal{F}_{\hat{S}_n} \right) \\
&= \mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}} \beta^t \big(h(t) - \Gamma(t)\big) + \mathcal{E}\left(\sum_{t=\hat{S}_{N-1}+1}^{\hat{S}_N} \beta^t \big(h(t) - \gamma(\hat{S}_{N-1})\big) \middle| \mathcal{F}_{\hat{S}_{N-1}} \right) \middle| \mathcal{F}_{\hat{S}_n} \right) \\
&= \mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\hat{S}_n} \right) = \mathcal{E}\left(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-2}} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\hat{S}_n} \right) \\
&= \cdots = 0.
\end{aligned}$$

By our definition of $\hat{S}_n$, we have $\hat{S}_N \geq N$. Hence, $\hat{S}_N \to \infty$ as $N \to \infty$. Therefore, by applying Lebesgue property, the result follows. □

Remark 8.3. Bank and El Karoui [11] consider a similar result to this proposition, but under a classical expectation with the summation $\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \gamma(s)\right)$ replaced by a more general function in continuous time. (See also Bank and Föllmer [12] and Bank and Küchler [13] for further discussion).

Intuitively, as $\Gamma(t) \geq \gamma(t-1)$, the process Γ should be sufficient to compensate for continuing to play. This means that the total 'expected' loss, evaluated from any point in time, must be non-positive if a reward $\Gamma(t)$ is offered. This is stated formally in the following lemma and theorem.

Lemma 8.8. *Let $\tau \in \mathcal{T}(s)$ with $1 \leq \tau \leq \sigma := \sigma(s, \lambda)$. Then*

$$\mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\sigma} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right) \leq 0.$$

Proof. Write $H_\tau := \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\sigma} \beta^t \big(h(t) - \lambda\big) \middle| \mathcal{F}_{s+\tau} \right)$ and $A := \{H_\tau > 0\}$.

Define $\tilde{\sigma} := \tau \, \mathbb{I}_A + \sigma \, \mathbb{I}_{A^c}$.

$$\begin{aligned}
\mathcal{E}\left(\sum_{t=s+1}^{s+\sigma} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right) &= \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) + \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\sigma} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right) \\
&\geq \sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \lambda\right) + \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\sigma} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right) \mathbb{I}_{A^c} \\
&= \mathcal{E}\left(\sum_{t=s+1}^{s+\tilde{\sigma}} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_{s+\tau} \right).
\end{aligned}$$

Moreover, the above inequality is strict on A. Hence, if A is not a $\mathbb{P}$-null set, it then follows from strict monotonicity that

$$\mathcal{E}\left(\sum_{t=s+1}^{s+\sigma} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_s \right) > \mathcal{E}\left(\sum_{t=s+1}^{s+\tilde{\sigma}} \beta^t \left(h(t) - \lambda\right) \middle| \mathcal{F}_s \right).$$

This contradicts the minimality of $\sigma(s, \lambda)$ established in Theorem 8.6. □

Theorem 8.9. *For all $N \in \mathbb{N}$,*

$$\mathcal{E}\left(\sum_{t=N+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_N \right) \leq 0.$$

Proof. Define $\tau_n := (\hat{S}_{n+1} \wedge N) \vee \hat{S}_n$. Since $\hat{S}_n$ is a stopping time for all $n \in \mathbb{N}$, so is τ_n. Hence, by Proposition 8.7 and Lemma 8.8,

$$\begin{aligned}
&\mathcal{E}\left(\sum_{t=\tau_n+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\tau_n} \right) \\
&\quad = \mathcal{E}\left(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}} \beta^t \big(h(t) - \Gamma(t)\big) + \mathcal{E}\left(\sum_{t=\hat{S}_{n+1}+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\hat{S}_{n+1}} \right) \middle| \mathcal{F}_{\tau_n} \right) \\
&\quad = \mathcal{E}\left(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\tau_n} \right) = \mathcal{E}\left(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}} \beta^t \big(h(t) - \gamma(\hat{S}_n)\big) \middle| \mathcal{F}_{\tau_n} \right) \leq 0.
\end{aligned}$$

Therefore, as $\{\hat{S}_n \leq N < \hat{S}_{n+1}\}$ is $\mathcal{F}_N$-measurable, by Lebesgue property and regularity (Remark 6.3),

$$\begin{aligned}
\mathcal{E}\left(\sum_{t=N+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_N \right) &= \mathcal{E}\left(\lim_{L\to\infty} \left(\sum_{n=0}^{L} \mathbb{I}_{\{\hat{S}_n \leq N < \hat{S}_{n+1}\}} \right) \left(\sum_{t=\tau_n+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \right) \middle| \mathcal{F}_N \right) \\
&= \lim_{L\to\infty} \sum_{n=0}^{L} \mathbb{I}_{\{\hat{S}_n \leq N < \hat{S}_{n+1}\}} \mathcal{E}\left(\sum_{t=\tau_n+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_N \right) \\
&= \lim_{L\to\infty} \sum_{n=0}^{L} \mathbb{I}_{\{\hat{S}_n \leq N < \hat{S}_{n+1}\}} \mathbb{I}_{\{\tau_n \geq N\}} \mathcal{E}\left(\sum_{t=\tau_n+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_N \right) \\
&= \lim_{L\to\infty} \sum_{n=0}^{L} \mathbb{I}_{\{\hat{S}_n \leq N < \hat{S}_{n+1}\}} \mathbb{I}_{\{\tau_n \geq N\}} \mathcal{E}\left(\mathcal{E}\left(\sum_{t=\tau_n+1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \middle| \mathcal{F}_{\tau_n} \right) \middle| \mathcal{F}_N \right) \leq 0
\end{aligned}$$

□

Remark 8.4. The above theorem says that, with compensation $\Gamma(t)$, at any point in time we expect to obtain a net reward from continuing to play, i.e. we have a non-positive expected total loss.

8.1.4 Step A.4: Reward Delay and Robust Representation Theorem

In Step A.3, we have shown that our reward Γ is defined to be (just) sufficient to encourage the player to continue playing (Theorem 8.9), i.e. the total expected loss

is zero, as in Proposition 8.7. We now show that taking a break from play cannot improve a player's expected discounted costs. We now formulate this observation by establishing the existence of a probability measure in our representing set $\mathcal{Q}$ such that the expected discounted costs, accounting for the break in play, have a lower bound close to zero. This result will be useful when considering multiple arms.

Theorem 8.10. *By Assumption 6.3, recall that $\mathcal{E}$ admits a robust representation of the form $\mathcal{E}(\,\cdot\,) = \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}(\cdot)$.*

For every fixed $\epsilon > 0$, there exists a probability measure $\mathbb{Q} \in \mathcal{Q}$ such that for every predictable decreasing process $(\alpha(t))_{t\geq 0}$ taking values in $[0,1]$, we have

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\infty} \alpha(t)\beta^t\big(h(t) - \Gamma(t)\big)\right) \geq -\epsilon.$$

Proof. By Proposition 8.7 and the robust representation theorem, for a fixed $\epsilon > 0$, we can find a probability measure $\mathbb{Q} \in \mathcal{Q}$ such that

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\right) \geq -\epsilon.$$

For each predictable decreasing process $(\alpha(t))_{t\geq 0}$ taking values in $[0,1]$, we define

$$\alpha^N(t) := \begin{cases} \alpha(t) & \text{for } t \leq N, \\ \alpha(N) & \text{for } t > N. \end{cases}$$

We claim that

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\infty} \alpha^N(t)\beta^t\big(h(t) - \Gamma(t)\big)\right) \geq -\epsilon. \tag{8.1.1}$$

Indeed, it is clear that the result holds when $N = 0$.

For the sake of induction, assume that the result holds for a given N. We then have

$$-\epsilon \leq \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{N} \alpha(t)\beta^t\big(h(t) - \Gamma(t)\big)\right) + \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=N+1}^{T} \alpha(N)\beta^t\big(h(t) - \Gamma(t)\big)\right). \tag{8.1.2}$$

By the robust representation theorem (Theorem 6.3),

$$\mathcal{E}(\,\cdot\,|\mathcal{F}_N) = \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}(\,\cdot\,|\mathcal{F}_N).$$

By Theorem 8.9, we know that

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=N+1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\middle|\mathcal{F}_N\right) \leq 0.$$

Since α is decreasing,

$$\big(\alpha(N) - \alpha(N+1)\big)\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=N+1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\middle|\mathcal{F}_N\right) \leq 0.$$

As α is predictable, by rearranging the above inequality, we obtain

$$\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty}\alpha(N)\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_N\bigg)\leq\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty}\alpha(N+1)\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_N\bigg).$$

Hence, by the tower property,

$$\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty}\alpha(N)\beta^t\big(h(t)-\Gamma(t)\big)\bigg)\leq\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty}\alpha(N+1)\beta^t\big(h(t)-\Gamma(t)\big)\bigg).$$

By substituting this into (8.1.2), we prove (8.1.1) with N replaced by $N+1$ and done the induction step.

By using the bounded convergence theorem, we can take $N\to\infty$ and obtain the required result. □

8.2 Part B: Analysis of multiple arms

We are now ready to consider the problem of choosing between multiple arms.

In Definition 7.4, we introduce our class of admissible control, which can be considered in our dynamic allocation problems. This class of control introduces a few natural ways of parameterising time. Therefore, we will use the following terminology to describe time evolution in different ways. This terminology will be useful in our discussion on the proof.

1. **'Play'** refers to the total number of (real) times that we play the system of bandit (i.e. playing multiple arms). (This corresponds to the time parameter of the simple form ρ which is briefly discussed earlier on in Remark 6.7 and later in Definition 7.5.)

2. **'Trial'** refers to the number of times that we play a specific arm. (This corresponds to the sum $\sum_{i=0}^{l}\tau_i^{(k)}$.)

3. **'Decision'** refers to the number of times that we make a decision between arms. (This corresponds to the time parameter for the choice process $(p_n)_{n\geq 0}$.)

4. **'Run'** refers to the number of times that we have made the decision to select a specific arm. (This corresponds to the time parameter of the allocation sequence $(\tau_i^{(k)})_{i\geq 0}$ for each fixed $k\in[K]$.)

Remark 8.5. The terms 'play' and 'trial' can be referred to without directly identifying the time allocation sequence. On the other hand, the terms 'decision' and 'run' need to be interpreted under a given time allocation sequence τ (Definition 7.2).

Remark 8.6. The kth component of the recording sequence η_n (Definition 7.3) represents the number of runs in the kth arm before the nth decision. We can see that the random variable η_n takes values in $\Big\{r\in\mathcal{S}:\sum_{k=1}^{K}r^{(k)}\leq n\Big\}$.

In order to prove C-optimality, we then consider the target function as briefly stated in (7.2.10).

Definition 8.5. For each $k \in [K]$, let $\big(h^{(k)}(t)\big)_{t\geq 1}$ be the uniformly bounded non-negative cost process at the tth trial of the kth arm with prevailing reward process $\big(\Gamma^{(k)}(t)\big)_{t\geq 1}$ (Definition 8.4). For an allocation strategy (τ, p), (Definition 7.4), we define the *Gittins' target function* by

$$V(\tau,p) := \mathfrak{E}\bigg(\sum_{n=1}^{\infty} \beta^n \Big(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\Big)\bigg)$$

where ρ is the simple form of (τ, p) with corresponding counting processes $t_n^\rho := \sum_{i=0}^{n-1} \mathbb{I}(\rho_i = \rho_{n-1})$, and $\mathfrak{E}$ is a partially consistent orthant nonlinear expectation, as in Definition 6.6.

Remark 8.7. We can also write $V(\tau, p)$ in terms of τ and p directly without identifying the simple form ρ. This is done in the proof of Theorem 8.13 in Step B.2. This definition, however, makes it clear that V depends on (τ, p) only through its simple form.

8.2.1 Step B.1: Fubini theorem and Suboptimality

In this subsection, we will show that considering generic stopping times and choice of bandits yields a non-negative expected loss. This can be shown using the robust representation result.

First, we recall the following corollary of Fubini's theorem.

Corollary 8.11. *Let $(G, \mathcal{G}, \mathbb{P})$ and $(H, \mathcal{H}, \mathbb{Q})$ be probability spaces. Let $\mathcal{G}'$ and $\mathcal{H}'$ be sub σ-algebras of $\mathcal{G}$ and $\mathcal{H}$ respectively, with $\mathcal{H}' := \{\varnothing, H\}$. Then, for any integrable random variable X on $(G \times H, \mathcal{G} \otimes \mathcal{H}, \mathbb{P} \otimes \mathbb{Q})$, we have*

$$\mathbb{E}^{\mathbb{P}\otimes\mathbb{Q}}\Big(X \,\Big|\, \mathcal{G}' \otimes \mathcal{H}'\Big) = \mathbb{E}^{\mathbb{P}}\Big(\int_H X(\,\cdot\,, h) d\mathbb{Q}(h) \,\Big|\, \mathcal{G}'\Big) \quad \mathbb{P}\otimes\mathbb{Q}\text{-a.s.}$$

Theorem 8.12. *For any allocation strategy (τ, p), we have*

$$V(\tau, p) \geq 0.$$

Proof. Let ρ be the simple form of (τ, p). Write $R_t^{(k)}$ for the total number of trials on other bandits before making the tth trial on the mth arm, i.e.

$$R_t^{(k)} := \sum_{i\neq k} \sum_{n=0}^{N_t^{(k)}} \mathbb{I}(\rho_n = i) \quad \text{where} \quad N_t^{(k)} := \inf\bigg\{N \geq 0 : \sum_{n=0}^{N} \mathbb{I}(\rho_n = k) = t\bigg\}.$$

Since $R_t^{(k)}$ does not depend on future realisations of the kth arm, $R_t^{(k)}$ is $\mathcal{F}_{t-1}^{(k)} \otimes \big(\bigotimes_{i=1, i\neq k}^{K} \mathcal{F}_\infty^{(i)}\big)$-measurable (taking the product in an appropriate order). Moreover, as $N_t^{(k)}$ is increasing in t, it follows that $R_t^{(k)}$ is increasing in t.

Now, fix $\epsilon > 0$. By Theorem 8.10, for each $k \in [K]$, we can find a probability measure $\mathbb{Q}^{(k)} \in \mathcal{Q}^{(k)}$ such that, for every adapted decreasing process $(\alpha^{(k)}(t))_{t\geq 0}$ taking values in $[0,1]$, we have

$$\mathbb{E}^{\mathbb{Q}^{(k)}}\left(\sum_{t=1}^{\infty}\alpha^{(k)}(t)\beta^t\big(h^{(k)}(t)-\Gamma^{(k)}(t)\big)\right)\geq -\epsilon. \tag{8.2.1}$$

Define

$$\tilde{\alpha}^{(k)}(t) := \int_{\prod_{i\neq k}\Omega^{(i)}} \beta^{R_t^{(k)}}\, d\Big(\bigotimes_{i\neq k}\mathbb{Q}^{(k)}\Big).$$

By Fubini's theorem, as $R_t^{(k)}$ is $\mathcal{F}_{t-1}^{(k)}\otimes\big(\bigotimes_{i=1,i\neq k}^{K}\mathcal{F}_{\infty}^{(i)}\big)$-measurable and $\beta\in(0,1]$, the process $\big(\tilde{\alpha}^{(k)}(t)\big)$ is an $\big(\mathcal{F}_t^{(k)}\big)$-predictable process taking values in $[0,1]$.

Moreover, $R_t^{(k)}$ is also $\mathcal{F}_{\infty}^{(k)}\otimes\big(\bigotimes_{i=1,i\neq k}^{K}\mathcal{F}_{\infty}^{(i)}\big)$-measurable, so by Corollary 8.11 we can write

$$\tilde{\alpha}^{(k)}(t)=\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\Big(\beta^{R_t^{(k)}}\,\Big|\,\tilde{\mathcal{F}}_{\infty}^{(k)}\Big)\quad\text{where}\quad\tilde{\mathcal{F}}_{\infty}^{(k)}:=\mathcal{F}_{\infty}^{(k)}\otimes\bigotimes_{i\neq k}^{M}\mathcal{F}_0^{(i)}.$$

As $t\mapsto R_t^{(k)}$ is increasing, it then follows that $\tilde{\alpha}^{(k)}(t)$ is decreasing in t. Hence, by Theorem 8.10 we obtain (8.2.1) with $\alpha^{(k)}$ replaced with $\tilde{\alpha}^{(k)}$.

By the definition of $\mathfrak{E}$ and Fubini's theorem, it follows that

$$\begin{aligned}
\mathfrak{E}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho})-\Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right)&\geq\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho})-\Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right)\\
&=\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\left(\sum_{k=1}^{K}\sum_{t=1}^{\infty}\beta^{R_t^{(k)}}\beta^t\left(h^{(k)}(t)-\Gamma^{(k)}(t)\right)\right)\\
&=\sum_{k=1}^{K}\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\left(\sum_{t=1}^{\infty}\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\left(\beta^{R_t^{(k)}}\beta^t\left(h^{(k)}(t)-\Gamma^{(k)}(t)\right)\Big|\tilde{\mathcal{F}}_{\infty}^{(k)}\right)\right)\\
&=\sum_{k=1}^{K}\mathbb{E}^{\bigotimes_{i=1}^{K}\mathbb{Q}^{(i)}}\left(\sum_{t=1}^{\infty}\tilde{\alpha}^{(k)}(t)\beta^t\left(h^{(k)}(t)-\Gamma^{(k)}(t)\right)\right)\\
&=\sum_{k=1}^{K}\mathbb{E}^{\mathbb{Q}^{(k)}}\left(\sum_{t=1}^{\infty}\tilde{\alpha}^{(k)}(t)\beta^t\left(h^{(k)}(t)-\Gamma^{(k)}(t)\right)\right)\quad\geq\quad-K\epsilon.
\end{aligned}$$

As ϵ is arbitrary, the result follows. □

8.2.2 Step B.2: Optimality

In this subsection, we will show that the strategy determined by a particular time allocation sequence yields a zero expected cost in the Gittins' target function.

Theorem 8.13. *For each $k\in[K]$, let $(\sigma_i^{(k)})_{i\geq 0}$ be the sequence of running maximum random times associated to the kth arm, as defined in Definition 8.3, i.e. we define $\sigma_i^{(k)}$ and $\hat{S}_i^{(k)}$ recursively by $S_i^{(k)}:=\sum_{l=0}^{i-1}\sigma_l^{(k)}$ and*

$$\sigma_i^{(k)}:=\inf\left\{\theta\geq 1:\gamma^{(k)}(S_i^{(k)}+\theta)>\gamma^{(k)}(S_i^{(k)})\right\}.$$

Then for any allocation strategy of the form (σ, p), we have

$$V(\sigma, p) \leq 0.$$

Proof. Recall the recording sequence η_n associated with (σ, p). We define the following notation, given an allocation sequence σ.

- $\tilde{\Theta}_n$ denotes the total number of plays of the system before making the nth decision. i.e.
$$\tilde{\Theta}_n := \sum_{k=1}^{K} \sum_{i=0}^{\eta_n^{(k)}-1} \sigma_i^{(k)}.$$
- $\tilde{\sigma}_n$ denotes the duration we decide to play following the nth decision time, i.e.
$$\tilde{\sigma}_n := \sigma_i^{(k)} \qquad \text{on the event } \{p_n = k, \eta_n^{(k)} = i\}.$$
N.B. $p_n = k$ means that we decide to play the kth arm at the nth decision. The event $\eta_n^{(k)} = i$ means that we have had i runs of the kth arm before the nth decision. Thus, we choose to make $\sigma_i^{(k)}$ more trials on this arm before making another decision.
- $\tilde{\Psi}_n^{(k)}$ denotes the total number of trials on the kth arm before making the nth decision, i.e.
$$\tilde{\Psi}_n^{(k)} := \sum_{i=0}^{i-1} \sigma_i^{(k)} \qquad \text{on the event } \{\eta_n^{(k)} = i\}.$$

Using this notation, we can define a variation on the Gittins' target fuction, with the restriction that we consider only the first N plays of the system, that is,

$$V(N, \sigma, p) := \mathfrak{E}\left(\sum_{n=0}^{N-1} \beta^{\tilde{\Theta}_n} \left(\sum_{l=1}^{\tilde{\sigma}_n} \beta^l \left(h^{(p_n)}(\tilde{\Psi}_n^{(p_n)} + l) - \Gamma^{(p_n)}(\tilde{\Psi}_n^{(p_n)} + l) \right) \right) \right)$$

with the convention $h^{(0)}(t) = \Gamma^{(0)}(t) = 0$ for all t.

By considering the simple form ρ of the strategy (σ, p) and applying Lebesgue property of $\mathfrak{E}$, we can show that V agrees with Definition 8.5 as $N \to \infty$, that is,

$$\lim_{N \to \infty} V(N, \sigma, p) = V(\sigma, p).$$

Hence, it suffices to show that $V(N, \sigma, p) \leq 0$ for all $N \in \mathbb{N}$. This will be proved by induction.

It is clear that $V(0, \sigma, p) = 0$. Fix $N \in \mathbb{N}$ and assume that $V(N, \sigma, p) \leq 0$. To show that $V(N+1, \sigma, p) \leq 0$, by subadditivity, it suffices to show that

$$\mathfrak{E}\left(\beta^{\tilde{\Theta}_N} \left(\sum_{l=1}^{\tilde{\sigma}_N} \beta^l \left(h^{(p_N)}(\tilde{\Psi}_N^{(p_N)} + l) - \Gamma^{(p_N)}(\tilde{\Psi}_N^{(p_N)} + l) \right) \right) \right) \leq 0.$$

Define the following random variables as in (7.2.6):

$$\Psi_r^{(k)} := \sum_{i=0}^{r^{(k)}-1} \sigma_i^{(k)} \qquad \text{and} \qquad \Theta_r := \sum_{k=1}^{K} \Psi_r^{(k)} \tag{8.2.2}$$

for $r \in \underline{\mathcal{S}}_N := \left\{ r \in \mathcal{S} : \sum_{k=1}^K r^{(k)} \le N, \ \right\}$.

Note that $\Psi_r^{(k)} = \tilde{\Psi}_N^{(k)}$ and $\Theta_r = \tilde{\Theta}_N$ on the event $\{\eta_N = r\}$. From the definition of the recording sequence (Definition 7.3), it follows that Θ_r and the event $A_N^{(r,k)} := \{\eta_N = r, p_N = k\}$ are both $\mathcal{F}(\Psi_r)$-measurable.

On an event $A_N^{(r,k)}$, $\tilde{\Psi}_N^{(k)} = \Psi_r^{(k)}$ and $\Psi_r^{(k)}$ is a stopping time with respect to the filtration $(\mathcal{F}_t^{(k)})_{t\ge 0}$. By considering the optimality obtained in Theorem 8.6, we can show that

$$
\begin{aligned}
&\mathfrak{E}\left(\beta^{\Theta_N} \left(\sum_{l=1}^{\tilde{\sigma}_N} \beta^l \left(h^{(p_N)}(\Psi_N^{(k)} + l) - \Gamma^{(p_N)}(\Psi_N^{(k)} + l) \right) \right) \right) \\
&\quad = \mathfrak{E}\left(\sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathbb{I}_{A_N^{(r,k)}} \beta^{\Theta_r} \left(\sum_{l=1}^{\sigma_r(k)} \beta^l \left(h^{(k)}(\Psi_r^{(k)} + l) - \Gamma^{(k)}(\Psi_r^{(k)} + l) \right) \right) \right) \\
&\quad = \mathfrak{E}\left(\sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathbb{I}_{A_N^{(r,k)}} \beta^{\Theta_r} \left(\sum_{l=1}^{\sigma_r(k)} \beta^l \left(h^{(k)}(\Psi_r^{(k)} + l) - \gamma^{(k)}(\Psi_r^{(k)}) \right) \right) \right) \\
&\quad \le \sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathfrak{E}\left(\mathbb{I}_{A_N^{(r,k)}} \beta^{\Theta_r} \left(\sum_{l=1}^{\sigma_r(k)} \beta^l \left(h^{(k)}(\Psi_r^{(k)} + l) - \gamma^{(k)}(\Psi_r^{(k)}) \right) \right) \right) \\
&\quad \le \sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathfrak{E}\left(\mathbb{I}_{A_N^{(r,k)}} \beta^{\Theta_r} \mathfrak{E}_{\Psi_r} \left(\sum_{l=1}^{\sigma_r(k)} \beta^l \left(h^{(k)}(\Psi_r^{(k)} + l) - \gamma^{(k)}(\Psi_r^{(k)}) \right) \right) \right) \\
&\quad = \sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathfrak{E}\left(\mathbb{I}_{A_N^{(r,k)}} \beta^{\Theta_r} \mathcal{E}^{(k)} \left(\sum_{l=1}^{\sigma_r(k)} \beta^l \left(h^{(k)}(\Psi_r^{(k)} + l) - \gamma^{(k)}(\Psi_r^{(k)}) \right) \middle| \mathcal{F}_{\Psi_r^{(k)}}^{(k)} \right) \right) \\
&\quad = \sum_{k=1}^K \sum_{r \in \underline{\mathcal{S}}_N} \mathfrak{E}(0) \quad = \quad 0.
\end{aligned}
$$

We see that $V(N+1, \sigma, p) \le 0$, and the desired result follows by induction. □

8.2.3 Step B.3: C-optimality

In the previous sections, we introduced an allocation problem when the prevailing process is offered as compensation (Definition 8.5). We also proved that the optimal value could be achieved by choosing a proper family of allocation time sequences (i.e., σ as in Theorem 8.13).

The prevailing reward process Γ for each arm is non-decreasing, and the optimal allocation sequences σ require us to make a new decision whenever the process Γ increases. By exploiting this fact, together with the discount effect, we will see that it is preferable to play the arm with the lowest value of Γ first. In particular, we can establish the Robust Gittins index theorem, which we repeat for the convenience of the reader.

Theorem 8.14 (Theorem 7.2: Robust Gittins theorem)**.** *Suppose that for each $m \in \mathcal{M}$. Let $\psi_n^{(m)}$ be the total number of trials of the mth bandit before the nth play of the system. i.e. $\psi_n^{(m)} := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k^* = m)$ (given an allocation strategy ρ^* up to time $n-1$).*

Then the allocation strategy ρ^* given (recursively) by

$$\rho_n^* \coloneqq \min\left\{m \in \mathcal{M} \;:\; m \in \arg\min_k \gamma^{(k)}(\psi_n^{(k)})\right\}$$

is C-optimal (Definition 6.9) under $\mathfrak{E}$ for the cost

$$g^\rho(n) = \beta^n h^{(\rho_{n-1})}(t_n^\rho) \quad \textit{where} \quad t_n^\rho = \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1}).$$

Proof. Recall the definition of $\Psi_r^{(m)}$ in (7.2.6). We can see that Ψ_r determines the orthant filtration when $\tilde{\eta}_n = r$ where $(\tilde{\eta}_n)$ is a recording sequence constructed from the time allocation sequence σ, i.e. when the mth bandit was run for $r^{(m)}$ times under the (optimal) allocation sequence σ. In particular, $\Psi_r^{(m)}$ corresponds to the number of trials on the mth bandit.

To explicitly define our choice sequence, we set

$$p_n^* \coloneqq \min\left\{m \in \mathcal{M} \;:\; m \in \arg\min_k \gamma^{(k)}(\Psi_r^{(k)})\right\} \quad \text{on the event} \quad \{\tilde{\eta}_n = r\}.$$

As $\Psi_r^{(m)}$ is an $(\mathcal{F}_t^{(m)})$-stopping time, $\gamma^{(m)}(\Psi_r^{(m)})$ is well-defined and is $\mathcal{F}(\Psi_r^{(m)})$-measurable. It also follows that

$$\begin{aligned}\left(\{\tilde{\eta}_n = r\} \cap \{p_n^* = m\}\right) = \{\tilde{\eta}_n = r\} &\cap \bigcap_{k=1}^{m} \left\{\gamma^{(m)}(\Psi_r^{(m)}) < \gamma^{(k)}(\Psi_r^{(k)})\right\} \\ &\cap \bigcap_{k=1}^{M} \left\{\gamma^{(k)}(\gamma^{(m)}(\Psi_r^{(m)}) \le \Psi_r^{(k)})\right\} \\ &\in \mathcal{F}(\Psi_r)\end{aligned}$$

Hence, p^* is a choice sequence for the allocation sequence σ (Definition 7.3). Therefore, (σ, p^*) is an admissible allocation strategy. Moreover, observe that ρ^* given in the statement of this theorem is the simple form of the allocation strategy (σ, p^*).

By Theorem 8.13 and Theorem 8.12,

$$\mathfrak{E}\left(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1}^*)}(t_n^*) - \Gamma^{(\rho_{n-1}^*)}(t_n^*)\right)\right) = 0 \quad \text{for} \quad t_n^* \coloneqq \sum_{k=0}^{n-1} \mathbb{I}(\rho_k^* = \rho_{n-1}^*).$$

Theorem 8.12 also implies that for any allocation strategy (τ, p) (and thus for any simple form ρ),

$$\mathfrak{E}\left(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\right)\right) \ge 0.$$

Next, we will show that $n \mapsto \beta^n \Gamma^{(\rho_{n-1})}(t_n^\rho)$ is predictable with respect to our observed filtration. We recall that

$$\Gamma^{(m)}(t) = \max_{0 \le \theta \le t-1} \gamma(\theta) \quad : \quad t = 1, 2, \ldots.$$

Now, observe that $t_n^\rho = \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1}) = 1+\eta_{n-1}^{(\rho_{n-1})}$ where (η_n) is a recording sequence corresponding to a strategy $(\mathtt{1}, \rho)$ and hence,

$$\Gamma^{(\rho_{n-1})}(t_n^\rho) = \sum_{r\in\mathcal{S}_{n-1}} \mathbb{I}(\eta_{n-1} = r)\mathbb{I}(\rho_{n-1} = m)\Gamma^{(m)}(1 + r^{(m)})$$

where $\mathcal{S}_N \coloneqq \left\{r \in \mathcal{S} \,:\, \sum_{m=1}^M r^{(m)} = N\right\}$.

For simplicity of our discussion to establish that our compensators are prediction, we assume that $(\mathcal{F}_t^{(m)})_{t\geq 0}$ is generated by some underlying process $(\xi_t^{(m)})_{t\geq 1}$. In fact, we can prove the predictability directly using the similar argument to Theorem B.8.

Since $\Gamma^{(m)}(1+r^{(m)})$ is $\mathcal{F}_{r^{(m)}}^{(m)}$-measurable, by the Doob–Dynkin lemma, there exists a measurable function $f_r^{(m)} : \mathbb{R}^{r^{(m)}} \to \mathbb{R}$ such that

$$\Gamma^{(m)}(1 + r^{(m)}) = f_r^{(m)}(\xi_1^{(m)}, ..., \xi_{r^{(m)}}^{(m)}). \tag{8.2.3}$$

Note that, while the process $\left(\xi_t^{(m)}\right)$ is defined on $(\Omega^{(m)}, \mathcal{F}^{(m)})$, we can extend it to $(\bar{\Omega}, \bar{\mathcal{F}})$ by considering an appropriate embedding.

By substituting into (8.2.3) in a similar way to (B.1.3), we can write $\Gamma^{(\rho_{n-1})}(t_n^\rho)$ as a (measurable) function of $\left((\eta_k)_{0\leq k\leq n-1}, (\rho_k)_{0\leq k\leq n-1}, \left(\xi_k^\rho\right)_{0\leq k\leq n-1}\right)$. By Definition 7.3 and Remark 7.6, (η_k) and (ρ_k) are adapted to the observed filtration $(\mathcal{H}_n^\rho)_{n\geq 0}$. It therefore follows that $\Gamma^{(\rho_{n-1})}(t_n^\rho)$ is $\mathcal{H}_{n-1}^\rho$-measurable.

Therefore, $C_0^\rho(n) \coloneqq \Gamma^{(\rho_{n-1})}(t_n^\rho)$ defines subcompensator at time $N = 0$ and it is fully compensated for $\rho = \rho^*$.

To construct compensator for a subsequent time N, we consider 'restarting' our system at an orthant time $r = (r^{(1)}, ..., r^{(M)}) \in \mathcal{S}_N \subseteq \mathcal{S}$ (as in Theorem B.8). As $\mathcal{F}(r)$ describes the information from all bandits, this needs to be done carefully. As each of our single-bandit filtrations $(\mathcal{F}_t^{(m)})_{t\geq 0}$ is generated by a discrete-time real-valued process, and $\mathcal{F}(r) = \bigotimes_m \mathcal{F}_{r^{(m)}}^{(m)}$, the Doob–Dynkin lemma states that any $\mathcal{F}(r)$-measurable random variable can be written as a Borel function of the first r observations. For concreteness, we denote these observations ω_r.

We proceed by freezing the value of ω_r and ω_u with $u \leq r$. Let $(\rho_n^{*,\omega_r})_{n\geq N}$ denote the minimum-Gittins-index strategy given by ρ^* defined in the theorem when we restart our analysis at r from a given ω_r. We do not change the Gittins indices γ when we fix ω_r, so the corresponding Γ processes satisfy

$$\Gamma_{\omega_r}^{(m)}(t) = \max_{r^{(m)}\leq\theta\leq t-1} \gamma^{(m)}(\theta) \leq \max_{u^{(m)}\leq\theta\leq t-1} \gamma^{(m)}(\theta) = \Gamma_{\omega_u}^{(m)}(t) \quad \text{for all } u \leq r$$

and are measurable with respect to ω_r. As discussed in Remark 7.2, the optimal strategy $(\rho_n^{*,\omega_r})_{n\geq N}$ coincides with the strategy ρ^* (and is therefore also measurable with respect to ω_r).

By repeating our earlier analysis, we see that, for each ω_r, as $\Gamma_{\omega_r}^{(\rho_{n-1}^{*,\omega_r})} \leq \Gamma_{\omega_u}^{(\rho_{n-1}^{*,\omega_r})}$, we can now unfreeze ω_r and ω_u, and summing over all possible scenario to show that

$$C_N^\rho(n) \coloneqq \beta^n \max_{\eta_N^{(\rho_{n-1})}\leq\theta\leq t_n^\rho-1} \gamma^{(\rho_n)}(\theta)$$

is decreasing in N. By applying the same argument as what have been done earlier, we also have $n \mapsto C_N^\rho(n)$ is $\mathcal{H}_n^\rho$-predictable. Therefore, $(C_N^\rho(n))$ defines subcompensator for strategy ρ and it fully compensates for the strategy ρ^*.

Finally, as $t \mapsto \Gamma_{\omega_r}^{(m)}(t)$ is increasing for all m and all ω_r, and ρ^* is a strategy where the lowest Γ is chosen first, it follows that for all $1 \le N \le \infty$, if $\rho \sim_N \rho^*$, then

$$\sum_{n=N+1}^{L} C_N^{\rho^*}(n) \le \sum_{n=N+1}^{L} C_N^{\rho}(n) \qquad \text{for all } \ L \ge N+1. \tag{8.2.4}$$

In particular, (6.4.5) is satisfied and therefore ρ^* is C-optimal. □

Chapter 9

Asymptotic behaviour of a Robust Gittins index

We have already discussed a few aspects of the multi-armed bandit as a stochastic control problem in discrete time.

We briefly mentioned in Section 2.1.2 for the computation of Gittins index that one may approximate the reward of a Gaussian bandit by using a continuous time argument [14, 112, 25, 21]. This analysis was studied independently from Gittins index theorem in continuous time [62, 13, 42, 78].

In this chapter, we will overview Gittins index theorem in continuous time and discuss the optimality and the allocation strategy related to the problem in Section 9.1. We then connect the theorem to the approximate result through the Kalman filter, including providing a few related examples in Section 9.2. In Section 9.3, we extend an argument of Bather [14] to provide an asymptotic approximation (i.e. an approximation when our observation size is large) to the robust Gittins problem for the Gaussian cost. This asymptotic approximation explains the phenomenon that the uncertainty aversion eventually dominates learning as observed in Figure 7.1.

9.1 Gittins theorem in continuous time

Let $(\Omega, \mathbb{P}, \mathcal{F})$ be a probability space and $(\mathcal{F}_t^{(k)})_{t\geq 0}$ be a filtration defined on this space for all $k \in [K]$. Here, $(\mathcal{F}_t^{(k)})_{t\geq 0}$ represents information from the kth arm as discussed in the earlier chapters[1]. We assume that if the kth arm is played for a small interval of time $[t, t+dt]$ where t is the amount of time that the kth arm has been played earlier, then we pay the cost $h^{(k)}(t)dt$.

Similar to his proposal in discrete time [77], Mandelbaum [78] also proposed the following definition to study the allocation strategy in continuous time.

Definition 9.1. A random process $\tilde{\eta}(t)$ taking values in $[0,\infty)^K$ is called a *dynamic allocation strategy* if it satisfies

[1] In this chapter, we no longer insist that each arm is defined on a separate space as in the earlier chapters, for simplicity of reference.

(i) For each $k \in [K]$, $t \mapsto \tilde{\eta}^{(k)}(t)$ is increasing and right continuous.

(ii) $\sum_{k=1}^{K} \tilde{\eta}^{(k)}(t) = t$ for all $t \in [0, \infty)$.

(iii) $\{\tilde{\eta}(t) \le r\} := \bigcap_{k=1}^{K} \{\tilde{\eta}^{(k)}(t) \le r^{(k)}\} \in \mathcal{F}(r) := \bigvee_{k=1}^{K} \mathcal{F}^{(k)}_{r^{(k)}}$.

It is easy to show that $t \mapsto \tilde{\eta}^{(k)}(t)$ is absolutely continuous with respect to Lebesgue measure. Therefore, without loss of generality, we may assume that

$$\tilde{\eta}^{(k)}(t) = \int_0^t U_s^{(k)} ds$$

where U is a process taking values in $\Delta^M := \{u \in [0,1]^M : \sum_{m=1}^{M} u^{(k)} = 1\}$.

The process U may be interpreted as our probabilistic decision strategy in discrete-time. We may also see this as a relaxed control as in Reisinger and Zhang [88] or in Wang et al. [107].

Using Mandelbaum's allocation strategy [78], El Karoui and Karatzas [42] also proved Gittins theorem in continuous time without requiring any Markov assumption.

Theorem 9.1 (El Karoui and Karatzas [42])**.** *We define the continuous time Gittins index to be*

$$\begin{aligned}\gamma^{(k)}(t) &:= \operatorname*{ess\,inf}_{\tau \in \mathcal{T}^{(k)}(t)} \frac{\mathbb{E}\big[\int_t^\tau e^{-\alpha s} h^{(k)}(s) ds \big| \mathcal{F}_t^{(k)}\big]}{\mathbb{E}\big[\int_t^\tau e^{-\alpha s} ds \big| \mathcal{F}_t^{(k)}\big]} \\ &= \operatorname{ess\,inf}\Big\{\gamma \in m\mathcal{F}_t^{(k)} : \operatorname*{ess\,inf}_{\tau \in \mathcal{T}^{(k)}(t)} \mathbb{E}\Big[\int_t^\tau e^{-\alpha s}(h^{(k)}(s) - \gamma) ds \Big| \mathcal{F}_t^{(k)}\Big] \le 0\Big\}.\end{aligned}$$

where $\mathcal{T}^{(k)}(t)$ is a family of stopping times $\tau > t$.

Then a dynamic allocation strategy $\tilde{\eta}^*(t) = \int_0^t U_s ds$ chosen such that

$$\{U_s^{(k)} = 1\} \qquad on \qquad \big\{\gamma^{(k)}(\tilde{\eta}^{*(k)}(t)) = \min \gamma^{(k)}(\tilde{\eta}^{*(k)}(t))\big\}$$

will minimise

$$V(\tilde{\eta}) := \mathbb{E}\Big[\int_0^\infty e^{-\alpha s}\Big(\sum_{m=1}^{M} h^{(k)}\big(\tilde{\eta}^{(k)}(t)\big) d\tilde{\eta}^{(k)}(t)\Big)\Big].$$

Remark 9.1. Bank and Küchler [13] also give an alternative proof with minimal assumptions of this result using a stochastic representation, derived in Bank and El Karoui [11]. Their proof removes the assumption on the quasi-left-continuity of the filtration from the original proof (in continuous time) of El Karoui and Karatzas [42]. They also find a necessary and sufficient condition to ensure that the strategy induced by Gittins index in continuous time is optimal.

9.1.1 Gittins index for Diffusion processes

Now we focus on the index of a single-arm and omit the superscript (k).

Suppose that the cost process $h(t)$ can be written as a function $h(X_t)$ where X_t is a diffusion process

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t$$

where B is a Brownian motion. This model is considered by Karatzas [62] where he proved the first version of Gittins' theorem in continuous time via the HJB equation. In addition to the proof, he also gives a semi-explicit formula for Gittins index for the one-dimensional diffusion process.

Proposition 9.2 (Karatzas [62]). *Let (X_t) be a one dimensional diffusion process and the cost process of a single arm can be written as a function $h(X_t)$ where h is monotone. Then Gittins index can be written as $\gamma(t) = \gamma(X_t)$ where*

$$\gamma(x) = \alpha\Big(\frac{p'(x)g(x) - g'(x)p(x)}{-g'(x)}\Big) = \alpha\bigg(p(x) - p'(x)\Big(\frac{g(x)}{g'(x)}\Big)\bigg)$$

where

$$p(x) = \mathbb{E}_x\Big(\int_0^\infty e^{-\alpha s}h(X_s)ds\Big) \quad \textit{and} \quad g(x) = \begin{cases} \mathbb{E}_x\Big(\exp(-\alpha\tau_0)\Big); & x > 0 \\ \Big[\mathbb{E}_0\Big(\exp(-\alpha\tau_x)\Big)\Big]^{-1}; & x \le 0. \end{cases}$$

Here, $\tau_y := \inf\{t \ge 0; X_t = y\}$.

Example 9.1 (Karatzas [62]). For $b(X_t) \equiv b$ and $\sigma(X_t) \equiv \sigma$, one can show that

$$\begin{aligned} g(x) &= \exp(-\gamma x) \\ p(x) &= \frac{2}{\sigma^2(\beta+\gamma)}\Big[e^{-\gamma x}\int_{-\infty}^{x} h(y)e^{\gamma y}dy + e^{\beta x}\int_x^\infty h(y)e^{-\beta y}dy\Big] \end{aligned}$$

where

$$\gamma = \frac{\sqrt{b^2 + 2\alpha\sigma^2} - b}{\sigma^2} \qquad \text{and} \qquad \beta = \frac{\sqrt{b^2 + 2\alpha\sigma^2} + b}{\sigma^2}.$$

It then follows that

$$\gamma(x) = \int_0^\infty h\Big(x + \frac{z}{\beta}\Big)e^{-z}dz.$$

9.2 Gittins' theorem with hidden state

We now connect Gittins' theorem in continuous time to the Kalman filtering.

9.2.1 Gittins index with hidden state

Let (Y_t) be an observation process taking values in $\mathbb{R}$. Suppose that the dynamic of (Y_t) depends on a hidden process $(\hat{X}_t)$ via

$$\begin{aligned} dY_t &= a_t\hat{X}_t dt + \varphi_t dW_t \quad ; & Y_0 &= 0 \\ d\hat{X}_t &= b_t\hat{X}_t dt + \psi_t d\hat{B}_t \quad ; & \hat{X}_0 &\sim N(X_0, p_0) \end{aligned}$$

where W and $\hat{B}$ are independent Brownian motions and the processes a_t, b_t, ψ_t and φ_t are deterministic with appropriate dimension.

Assume that when the arm is played for a small interval of time $[t, t+dt]$, we pay the cost dY_t. Here, t is the amount of time previously spent on this arm.

As we observe the process (Y_t), by Kalman filtering theory (see e.g. Cohen and Elliot [33]), we have

$$\left.\begin{aligned} dX_t &= b_t X_t dt + k_t \varphi_t dB_t, \\ dp_t/dt &= b_t p_t + p_t b_t^T + \psi_t \psi_t^T - k_t \varphi_t \varphi_t^T k_t^T, \\ dY_t &= a_t X_t dt + \varphi_t dB_t, \end{aligned}\right\} \tag{9.2.1}$$

where $k_t := p_t a_t^T \big(\varphi_t \varphi_t^T\big)^{-1}$ is the *Kalman gain process*. The process B_t is called the *innovations process* which is a Brownian motion under $(\mathcal{F}_t^Y)$.

In particular, if we denote $Y_t^{(k)}$ as the continuous time payoff of the kth arm, the objective of the multi-armed bandit problem in the Kalman filtering setting is to find a dynamic allocation strategy (Definition 9.1) $\tilde{\eta}^*$ to minimise the cumulative cost

$$\text{“ } V(\tilde{\eta}) := \mathbb{E}\Big[\int_0^\infty e^{-\alpha t} \Big(\sum_{k=1}^K dY^{(k)}_{\tilde{\eta}^{(k)}(t)} \Big) \Big]. \text{ ”}$$

It then follows from the martingale property that

$$V(\tilde{\eta}) = \mathbb{E}\Big[\int_0^\infty e^{-\alpha t} \Big(\sum_{k=1}^K a_{\tilde{\eta}^{(k)}(t)} X_{\tilde{\eta}^{(k)}(t)} \Big) d\tilde{\eta}^{(k)}(t) \Big].$$

In particular, the Kalman-filter model can be simplified to El Karoui and Karatzas' continuous time model (Theorem 9.1) with $h(t) = a_t X_t$.

Mean-reversing Gittins

Suppose that $a_t \equiv a$, $b_t \equiv -\lambda$, $\varphi_t \equiv \varphi$ and $\psi_t \equiv \psi$ in (9.2.1) are all scalar with $\lambda, \varphi > 0$. Suppose that $p_0 = p^*$ where p^* is the positive root of the Ricatti equation

$$0 = -2\lambda p^* + \psi^2 - (a^2/\varphi^2) p^{*2}$$

We can write

$$dX_t = -\lambda X_t dt + \sigma dB_t \qquad \text{where} \qquad \sigma := a p^*/\varphi.$$

In particular, we have (X_t) as an OU process with parameter λ and σ. We can now apply Proposition 9.2 to establish an explicit formula for Gittins' index in this case using an explicit formula for the Laplace Transform of an OU process.

Proposition 9.3 (Alili et al. [3] or Breiman [20])**.** *Let x be an initial point of the OU process. Then for $a > x$,*

$$\mathbb{E}_x \exp(-\alpha \tau_a) = \frac{H_{-\alpha/\lambda}(-x\sqrt{\lambda}/\sigma)}{H_{-\alpha/\lambda}(-a\sqrt{\lambda}/\sigma)}$$

where H_ν is the Hermite function (Lebedev [73, Chapter 10]) given by

$$H_\nu(z) := \frac{1}{2\Gamma(-\nu)} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \Gamma\Big(\frac{m-\nu}{2}\Big)(2z)^m.$$

Moreover, H_ν satisfies the recurrence relations

$$\frac{d}{dz} H_\nu(z) = 2\nu H_{\nu-1}(z) \qquad \textit{and} \qquad H_{\nu+1}(z) = 2zH_\nu(z) - 2\nu H_{\nu-1}(z).$$

We can now use the results from Karatzas [62] and Alili et al. [3] to establish Gittins' index for a one-dimensional Kalman Filtering model at equilibrium (ergodic limit).

In our case of interest, for $x > 0$, we have

$$\mathbb{E}_x \exp(-r\tau_0) = \frac{H_{-\alpha/\lambda}(x\sqrt{\lambda}/\sigma)}{H_{-\alpha/\lambda}(0)}.$$

On the other hand, for $x \le 0$,

$$\mathbb{E}_0 \exp(-\alpha\tau_x) = \frac{H_{-\alpha/\lambda}(0)}{H_{-\alpha/\lambda}(x\sqrt{\lambda}/\sigma)}.$$

Hence, in the notation of Proposition 9.2, we have

$$g(x) = \frac{H_{-\alpha/\lambda}(x\sqrt{\lambda}/\sigma)}{H_{-\alpha/\lambda}(0)}.$$

We recall that the OU process starting at x can be expressed explicitly by

$$X_t = e^{-\lambda t}\Big(x + \sigma \int_0^t \exp(\lambda s) dB_s\Big).$$

It follows from integration by parts that, in the notion of Proposition 9.2, we have

$$\begin{aligned} p(x) &= a\mathbb{E}_x\Big(\int_0^\infty e^{-\alpha s} X_s ds\Big) = a\mathbb{E}_x\Big(\Big[-e^{-\alpha s}X_s/\alpha\Big]_0^\infty + \frac{1}{\alpha}\int_0^\infty e^{-\alpha s} dX_s\Big) \\ &= \frac{ax}{\alpha} + a\mathbb{E}_x\Big(-\frac{\lambda}{\alpha}\int_0^\infty e^{-\alpha s} X_s ds + \frac{1}{\alpha}\int_0^\infty \sigma e^{-\alpha s} dB_s\Big) = \frac{ax}{\alpha} - \frac{\lambda}{\alpha} p(x) = \frac{ax}{\alpha+\lambda}. \end{aligned}$$

Now, we write $\nu = -\alpha/\lambda$. It follows from Proposition 9.2 and the recurrence property of H_ν that

$$\begin{aligned} \gamma(x) &= \alpha\left(p(x) - p'(x)\left(\frac{g(x)}{g'(x)}\right)\right) = \alpha\left(\left(\frac{ax}{\alpha+\lambda}\right) - \frac{1}{2}\left(\frac{a}{\alpha+\lambda}\right)\left(\frac{H_\nu(x\sqrt{\lambda}/\sigma)}{\nu(\sqrt{\lambda}/\sigma)H_{\nu-1}(x\sqrt{\lambda}/\sigma)}\right)\right) \\ &= \alpha\left(\frac{ax}{\alpha+\lambda}\right) + \frac{1}{2}\left(\frac{a\sigma\sqrt{\lambda}}{\alpha+\lambda}\right)\left(\frac{H_\nu(x\sqrt{\lambda}/\sigma)}{H_{\nu-1}(x\sqrt{\lambda}/\sigma)}\right)\Big). \end{aligned}$$

Gittins index for the Gaussian cost process

Previously, we consider when the underlying state $(\hat{X}_t)$ is ergodic. We now consider when $(\hat{X}_t)$ is stationary.

Suppose that $a_t \equiv 1$, $b_t, \psi_t \equiv 0$ and $\varphi_t \equiv 1$ in (9.2.1). In particular, we consider a model where the cost of an arm is driven by an unknown parameter Θ. The corresponding cost process is given by

$$dY_t = \Theta dt + dW_t$$

where W is a Brownian motion.

Remark 9.2. One may see this model as a continuous time analogy to the classical multi-armed bandit problem where the reward of the kth arm is generated from a distribution $N(\Theta^{(k)}, 1)$ where $\Theta^{(k)}$ is unknown. The parameters $\Theta^{(k)}$'s are chosen independently from a normal prior.

By the Kalman Filter, we have $\Theta \sim N(X_t, p_t)$,

$$dX_t = p_t dB_t, \qquad dp_t = -p_t^2 dt, \qquad \text{and} \quad dY_t = X_t dt + dB_t.$$

where B is a Brownian motion with respect to the observed filtration and $p_0 > 0$ is an initial prior variance.

By Theorem 9.1, Gittins index can be given by

$$\gamma(t) = \operatorname{ess\,inf}\Big\{\gamma \in m\mathcal{F}_t : \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(t)} \mathbb{E}\Big[\int_t^\tau e^{-\alpha s}(X_s - \gamma)ds \Big| \mathcal{F}_t\Big] \le 0\Big\}.$$

Since a pair (X_t, p_t) is a Markov process, we can write $\gamma(t) = \gamma(X_t, p_t)$ where

$$\gamma(x,p) = \inf\Big\{\gamma \in \mathbb{R} : \inf_{\tau > 0} \mathbb{E}\Big[\int_0^\tau e^{-\alpha s}(X_s - \gamma)ds \Big| X_0 = x, p_0 = p\Big] \le 0\Big\}.$$

By continuity of the optimal stopping time, we can show that $\gamma(x,p)$ is the minimum value of γ such that

$$0 = \inf_{\tau \ge 0} \mathbb{E}\Big[\int_0^\tau e^{-\alpha s}(X_s - \gamma)ds \Big| X_0 = x, p_0 = p\Big]. \tag{9.2.2}$$

By using integration by parts, we obtain

$$0 = \inf_{\tau \ge 0} \mathbb{E}\Big[(\gamma - X_\tau)e^{-\alpha\tau} - (\gamma - X_0) - \int_0^\tau e^{-\alpha s} dX_s \Big| X_0 = x, p_0 = p\Big].$$

As X is a martingale, so is $\left(\int_0^t e^{-\alpha s} dX_s\right)_{t \ge 0}$. It then follows from Doob's optional stopping theorem that

$$0 = \inf_{\tau \ge 0} \mathbb{E}\Big[(\gamma - X_\tau)e^{-\alpha\tau} - (\gamma - X_0) \Big| X_0 = x, p_0 = p\Big].$$

Define $Z_t := X_t - \gamma(x,p)$ and $z = x - \gamma(x,p)$.

By rearranging the formula above, we obtain

$$z = \sup_{\tau \geq 0} \mathbb{E}\big[Z_\tau e^{-\alpha\tau} \big| Z_0 = z, p_0 = p \big].$$

As $\gamma(x,p)$ is the minimal value of γ satisfying (9.2.2), it follows that (z,p) must be a free boundary of the optimal stopping problem

$$v(z,p) = \sup_{\tau \geq 0} \mathbb{E}\big[Z_\tau e^{-\alpha\tau} \big| Z_0 = z, p_0 = p \big].$$

By the Hamilton–Jacobi–Bellman equation (see e.g. Pham [84]), the corresponding variational inequality can be given by

$$\begin{aligned} \frac{1}{2} p^2 \, \partial_{zz} v - p^2 \, \partial_p v - \alpha v \begin{cases} = 0 & : (z,p) \in \mathcal{C} \\ \leq 0 & : (z,p) \notin \mathcal{C} \end{cases} \\ v \begin{cases} > z & : (z,p) \in \mathcal{C} \\ = z & : (z,p) \notin \mathcal{C} \end{cases} \end{aligned} \tag{9.2.3}$$

where $\mathcal{C}$ is a continuation region of the optimal stopping problem for (z,p).

We now introduce the following change of variables:

$$s = p/\alpha, \qquad w = z/\sqrt{\alpha}, \qquad \text{and} \qquad u = e^{-1/s} v/\sqrt{\alpha}.$$

Then the free-boundary (9.2.3) becomes

$$\begin{aligned} \frac{1}{2} \, \partial_{ww} u - \partial_s u \begin{cases} = 0 & : (w,s) \in \mathcal{C} \\ \leq 0 & : (w,s) \notin \mathcal{C} \end{cases} \\ u \begin{cases} > w \exp(-1/s) & : (w,s) \in \mathcal{C} \\ = w \exp(-1/s) & : (w,s) \notin \mathcal{C} \end{cases} \end{aligned} \tag{9.2.4}$$

where $\mathcal{C}$ is a continuation region of the optimal stopping problem for (w,s).

To find the free-boundary of (9.2.4), we will first show that the free-boundary of (9.2.4) has a monotonicity property.

Theorem 9.4 (Bather [14], Chang and Lai [25])**.** *Consider the free-boundary problem*

$$\begin{aligned} \frac{1}{2} \, \partial_{ww} u - \partial_s u \begin{cases} = 0 & : (w,s) \in \mathcal{C} \\ < 0 & : (w,s) \notin \mathcal{C} \end{cases} \\ u \begin{cases} \geq R(w,s) & : (w,s) \in \mathcal{C} \\ = R(w,s) & : (w,s) \notin \mathcal{C} \end{cases} \end{aligned} \tag{9.2.5}$$

Then the free-boundary problem (9.2.5) *corresponds to the optimal stopping problem:*

$$u(w,s) = \sup_{\tau \leq s} \mathbb{E}\Big[R(w + W_\tau, s - \tau) \Big]$$

where W is a standard Brownian motion and the region $\mathcal{C}$ corresponding to the continuation region of the optimal stopping problem.

Furthermore, if we have $R(w,s) = w\exp(-1/s) + L(s)$ where $L(s)$ is a decreasing function, then there exists a non-negative function $w^*(s)$ such that

$$\mathcal{C} = \{(w,s) : w < w^*(s)\}. \tag{9.2.6}$$

Proof. The stated theorem is almost identical to Bather [14] and Chang and Lai [25] but we add a slight modification in the theorem for our future reference. Nonetheless, the proof follows in the same spirit. □

The above result allows us to characterise $\mathcal{C}$ by the function $s \mapsto w^*(s)$[2]. We may then approximate $\mathcal{C}$ by considering the asymptotic behaviour of w^*.

Theorem 9.5 (Bather [14] and Chang and Lai [25])**.** *The free-boundary $w^*(s)$ for $R(w,s) = w\exp(-1/s)$ satisfies*

(i) $s \mapsto w^*(s)/\sqrt{s}$ *is increasing.*

(ii) As $s \to 0$, $w^*(s)/s \to 1/\sqrt{2}$.

(iii) As $s \to \infty$, $w^*(s) = s^{1/2}\sqrt{\log s - \frac{1}{2}\log\log s - \frac{1}{2}\log 16\pi + o(1)}$.

By using the corrected binomial method of Chernoff and Petkau [27], Brezzi and Lai [21] find a numerical solution to the free-boundary $w^*(s)$. Combining this with the asymptotic result (Theorem 9.5), Brezzi and Lai [21] proposed a simple closed form expression to find $w^*(s)$ by $w^*(s) \approx \sqrt{s}\psi(s)$ where

$$\psi(s) = \begin{cases} \sqrt{s/2} & : s \le 0.2 \\ 0.49 - 0.11s^{-1/2} & : 0.2 < s \le 1 \\ 0.63 - 0.26s^{-1/2} & : 1 < s \le 5 \\ 0.77 - 0.58s^{-1/2} & : 5 < s \le 15 \\ \left(2\log s - \log\log s - \log 16\pi\right)^{1/2} & : s \ge 15. \end{cases}$$

Therefore, the Gittins index in this case can be approximated by

$$\gamma(x,p) \approx x - \sqrt{\alpha}w^*(s) = x - \sqrt{p}\psi(p/\alpha).$$

This result is similar to an asymptotic expansion for the discrete model considered in Russo [95] where he argues that if x and p are posterior mean and variance of the cost, then

$$\gamma(x,p) = x - \sqrt{p}\Phi^{-1}(\beta) + o(1) \qquad \text{as} \qquad \beta \to 1$$

where β is a discount factor in discrete time, and Φ is the cumulative density function of the standard normal distribution.

Remark 9.3. Throughout the above discussion, we consider the problem in terms of cost minimisation to avoid confusion when extending those results to the robust index framework. The reader may find the results discussed in this section differ from the original material by an appropriate sign change.

[2]The existence of the representation (9.2.6) can be equivalently interpreted as this arm is Whittle indexable (see section 2.1.3)

9.3 Asymptotic behaviour of a Robust Gittins index

We can now give an approximation to the asymptotic behaviour of the robust Gittins' index discussed in chapter 6, 7 and 8 when we have a Gaussian cost.

We will follow the same argument as in Bather [14] to show that when t is large, the robust index can be decomposed into two components which correspond to the learning and uncertainty aversion. In particular, we will show that as $t \to \infty$

$$\gamma(x,t) \approx x + k\sigma\Big(\frac{1}{\sqrt{t}}\Big) - \frac{\sigma}{\sqrt{2\alpha}}\Big(\frac{1}{t}\Big). \tag{9.3.1}$$

Here, x is the estimated reward, t is a proxy for the number of observation, k is the size of confidence width which is used to construct nonlinear expectation, σ is the volatility of the observed signal and α is a discount rate in continuous time (which shall be interpreted as $1-\beta$ in the discrete time).

The term $k\sigma(\frac{1}{\sqrt{t}})$ corresponds to uncertainty aversion whereas the term $\frac{\sigma}{\sqrt{2\alpha}}(\frac{1}{t})$ explains an asymptotic value for learning. We can see in the expression (9.3.1) that when t is large, the uncertainty aversion dominates the learning term. This explains the domination of uncertainty aversion that we observed in Figure 7.1 when n is large and β is small.

For the rest of this chapter, we will derive an approximation (9.3.1). We first discuss how to define a nonlinear expectation in continuous time to model uncertainty aversion using upper confidence bound as in DR-Expectation of Cohen [29], and Bielecki et al. [16].

9.3.1 Data-Robust Nonlinear Expectation in continuous time

Let $(\Omega, \mathbb{P}, \mathcal{F})$ be an underlying probability space equipped with a process (Y_t) starting from y_0 at time $t_0 > 0$ such that Y/σ is a Brownian motion under $\mathbb{P}$ for some $\sigma > 0$. We will regard Y as our observed cost process as in the Kalman filter framework (Section 9.2.1). We define the filtration $(\mathcal{F}_t)_{t\geq t_0}$ to be an augmented filtration generated by the process Y.

Under $\mathbb{P}$, we can write $dY_t = \sigma dW_t$ where W is a Brownian motion.

Suppose that under the true probability $\mathbb{P}_{\mu^*}$, (Y_t) follows the dynamic

$$dY_t = \mu^* dt + \sigma dW_t^{\mu^*}$$

where W^{μ^*} is a Brownian motion under $\mathbb{P}_{\mu^*}$. We will assume that the value μ^* is not known. Hence, at time t, we need do an inference for μ^* from a confidence interval $[X_t - k\sigma/\sqrt{t}, X_t + k\sigma/\sqrt{t}]$ where $X_t := Y_t/t$ which is well-defined for $t > t_0$.

To model uncertainty, we consider a family of probability $\mathcal{Q} := \{\mathbb{P}_\mu : \mathbb{P}_\mu << \mathbb{P}\}$ such that under $\mathbb{P}_\mu$

$$dY_t = \mu_t dt + \sigma dW_t^{\mu}$$

where W^μ is an $(\mathcal{F}_t)$-Brownian motion under $\mathbb{P}_\mu$ and $\mu_t \in [X_t - k\sigma/\sqrt{t}, X_t + k\sigma/\sqrt{t}]$, i.e., we consider μ_t as our recognised value of μ^* at time t.

We obtain the following result which follows from a standard application of the Backward Stochastic Differential Equation (BSDE).

Theorem 9.6. *An operation* $\mathcal{E}\big(\cdot\big|\mathcal{F}_t\big) \coloneqq \operatorname{ess\,sup}_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\big(\cdot\big|\mathcal{F}_t\big) : t \geq t_0$ *defines a* $(\mathcal{F}_t)_{t\geq t_0}$*-consistent coherent nonlinear expectation.*

In Definition 7.1, the robust Gittins index associated to the discrete-time cost process $(h(s))_{s\in\mathbb{N}}$ is given by

$$\gamma(t) \coloneqq \operatorname{ess\,inf}\left\{\gamma \in L^\infty(\mathcal{F}_t) : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(t)} \mathcal{E}\left(\sum_{s=t+1}^{\tau} \beta^t\big(h(s)-\gamma\big) \,\middle|\, \mathcal{F}_t\right) \leq 0\right\},$$

where $\beta \in (0,1)$ is a discount factor, $\mathcal{E}$ is a $(\mathcal{F}_s)$-consistent coherent nonlinear expectation and $\mathcal{T}(t)$ is the family of $\big(\mathcal{F}_s\big)$-bounded stopping times, $\tau > t$.

It is natural to approximate the robust Gittins index by considering the continuous version of the nonlinear optimal stopping problem above. For $t \geq t_0$, we define

$$\gamma(t) \coloneqq \operatorname{ess\,inf}\left\{\gamma \in L^\infty(\mathcal{F}_t) : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(t)} \operatorname*{ess\,sup}_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\left(\int_t^\tau e^{-\alpha s}\big(dY_s - \gamma ds\big) \,\middle|\, \mathcal{F}_t\right) \leq 0\right\}.$$

where $\mathcal{T}(t)$ is a family of stopping $\tau > t$.

By using a similar argument to Theorem 8.1, we can establish the following lemma.

Lemma 9.7. $\gamma \coloneqq \gamma(y_0, t_0)$ *is the unique minimal real solution to*

$$\begin{aligned} 0 &= \inf_{\tau\in\mathcal{T}(t_0)} \sup_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\Big(\int_{t_0}^\tau e^{-\alpha s}\big(dY_s - \gamma ds\big)\Big|Y_0 = y_0\Big)\Big) \\ &= \sup_{\tau\in\mathcal{T}(t_0)} \inf_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\Big(\int_{t_0}^\tau e^{-\alpha s}(\gamma - \mu_s)ds\Big|Y_0 = y_0\Big). \end{aligned}$$

9.3.2 Optimal Stopping problem and the corresponding free-boundary

In this section, we will argue that the sup inf problem in Lemma 9.7 is related to a classical optimal stopping.

We may represent a process μ by

$$\mu_s = X_s + \Big(\frac{k\sigma}{\sqrt{s}}\Big)U_s$$

where $X_s = Y_s/s$ and U_s is a control process taking values in $[-1,1]$.

By Itô's formula,

$$dX_s = \frac{1}{s}dY_s - \frac{Y_s}{s^2}ds = -\frac{1}{s}X_s ds + \frac{1}{s}\Big(\mu_s ds + \sigma dW_s^\mu\Big) = k\sigma s^{-3/2}U_s ds + \sigma s^{-1}dW_s^\mu.$$

For a fixed τ,

$$\begin{aligned} \inf_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\Big[\int_{t_0}^\tau e^{-\alpha s}(\gamma-\mu_s)ds\Big|Y_0 = y_0\Big] &= \inf_U \mathbb{E}_\mu\Big[\int_{t_0}^\tau e^{-\alpha s}\Big(\gamma - X_s - \Big(\frac{k\sigma}{\sqrt{s}}\Big)U_s\Big)ds\Big|Y_0 = y_0\Big] \\ &= \inf_U \mathbb{E}_\mu\Big[\int_{t_0}^\tau e^{-\alpha s}\Big(\gamma - x_{t_0} - \int_{t_0}^s k\sigma s^{-3/2}U_t dt - \int_{t_0}^s \sigma t^{-1}dW_t^\mu - \Big(\frac{k\sigma}{\sqrt{s}}\Big)U_s\Big)ds\Big|Y_0 = y_0\Big]. \end{aligned}$$

As W^μ is a Brownian motion under $\mathbb{P}_\mu$, we can see the control problem above as if the probability measure is independent of U. We can then see that the expression above is minimised when $U_s \equiv 1$. Therefore,

$$\sup_{\tau\in\mathcal{T}(t_0)} \inf_{\mathbb{P}_\mu\in\mathcal{Q}} \mathbb{E}_\mu\Big(\int_{t_0}^{\tau} e^{-\alpha s}(\gamma - \mu_s)ds \Big| Y_0 = y_0\Big) = \sup_{\tau\in\mathcal{T}(t_0)} \mathbb{E}\Big(\int_{t_0}^{\tau} e^{-\alpha s}(\gamma - X_s - k\sigma s^{-1/2})ds \Big| Y_0 = y_0\Big)$$

where

$$dX_s = k\sigma s^{-3/2}ds + \sigma s^{-1}dW_s$$

and W is a Brownian motion under $\mathbb{P}$.

By considering Lemma 9.7, we may estimate $\gamma(t_0)$ by the minimal γ such that

$$0 = \sup_{\tau>t_0} \mathbb{E}\Big(\int_{t_0}^{\tau} e^{-\alpha(s-t_0)}(\gamma - X_s - k\sigma s^{-1/2})ds \Big| X_{t_0} = x_0\Big). \tag{9.3.2}$$

Here, τ is a an $(\mathcal{F}_t)$-stopping time such that $\tau > t_0$.

We will introduce a change of variable to convert the problem (9.3.2) to a free-boundary problem as in (9.2.3).

Define $Z_t = (X_t + 2k\sigma t^{-1/2} - \gamma)/\sigma$ and $z_0 = (x_0 + 2k\sigma t_0^{-1/2} - \gamma)/\sigma$.

The equation (9.3.2) becomes

$$0 = \sup_{\tau>t_0} \mathbb{E}\Big(\int_{s=t_0}^{\tau} e^{-\alpha s}\big(k\sigma s^{-1/2} - \sigma Z_s\big)ds \Big| Z_{t_0} = z_0\Big)$$

where

$$dZ_t = t^{-1}dW_t.$$

Observe that

$$\int_{s=t_0}^{\tau} \big(- Z_s\big)e^{-\alpha s}ds = \frac{1}{\alpha}Z_\tau e^{-\alpha\tau} - \frac{1}{\alpha}Z_{t_0}e^{-\alpha t_0} - \frac{1}{\alpha}\int_{s=t_0}^{\tau} e^{-\alpha s}t^{-1}dW_s$$

Therefore, by using an integration by parts, we obtain

$$\begin{aligned} 0 &= \sup_{\tau>t_0} \mathbb{E}\Big(\frac{1}{\alpha}Z_\tau e^{-\alpha\tau} - \frac{1}{\alpha}z_0 e^{-\alpha t_0} + \int_{s=t_0}^{\tau} e^{-\alpha s}k\sigma s^{-1/2}ds \Big| Z_{t_0} = z_0\Big) \\ z_0 e^{-\alpha t_0} &= \sup_{\tau>t_0} \mathbb{E}\Big(Z_\tau e^{-\alpha\tau} + \alpha k\int_{s=t_0}^{\tau} e^{-\alpha s}s^{-1/2}ds \Big| Z_{t_0} = z_0\Big) \end{aligned}$$

Denoting $\Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-\theta^2)d\theta$. We obtain

$$z_0 e^{-\alpha t_0} + k\sqrt{\alpha\pi}\,\Phi(\sqrt{\alpha t_0}) = \sup_{\tau>t_0} \mathbb{E}\Big(Z_\tau e^{-\alpha\tau} + k\sqrt{\alpha\pi}\,\Phi(\sqrt{\alpha\tau}) \Big| Z_{t_0} = z_0\Big). \tag{9.3.3}$$

We may interpret the RHS as the continuous value for the optimal stopping problem

$$g(z,t) = \sup_{\tau\geq t} \mathbb{E}\Big(Z_\tau e^{-\alpha\tau} + k\sqrt{\alpha\pi}\,\Phi(\sqrt{\alpha\tau}) \Big| Z_t = z\Big). \tag{9.3.4}$$

One can see in the equation (9.3.3) that at a point (z_0, t_0), it is indifferent to continue or stop. Hence, (z_0, t_0) must lie at the boundary of the optimal stopping problem (9.3.4).

We can also write a variational inequality corresponding to the optimal stopping problem (9.3.4) as followed.

$$\begin{aligned} \frac{1}{2}t^{-2}\,\partial_{zz}\,g + \partial_t\,g \begin{cases} = 0 & : (z,t) \in \mathcal{C} \\ \leq 0 & : (z,t) \notin \mathcal{C} \end{cases} \\ g(z,t) \begin{cases} > ze^{-\alpha t} + k\sqrt{\alpha\pi}\,\Phi(\sqrt{\alpha t}) & : (z,t) \in \mathcal{C} \\ = ze^{-\alpha t} + k\sqrt{\alpha\pi}\,\Phi(\sqrt{\alpha t}) & : (z,t) \notin \mathcal{C}. \end{cases} \end{aligned} \tag{9.3.5}$$

We again introduce another change of variables to obtain a classical free-boundary problem as considered in Bather [14].

We introduce the following changes of variables.

$$s = \frac{1}{\alpha t}, \qquad w = \frac{z}{\sqrt{\alpha}} \qquad \text{and} \qquad h(w,s) = \sqrt{\alpha} g(z,t) - k\sqrt{\pi}$$

We then obtain a standard free boundary problem as considered in Theorem 9.4.

$$\begin{aligned} \frac{1}{2}\,\partial_{ww}\,h - \partial_s\,h \begin{cases} = 0 & : (w,s) \in \mathcal{C} \\ \leq 0 & : (w,s) \notin \mathcal{C} \end{cases} \\ h(w,s) \begin{cases} > R(w,s) & : (w,s) \in \mathcal{C} \\ = R(w,s) & : (w,s) \notin \mathcal{C} \end{cases} \end{aligned} \tag{9.3.6}$$

where

$$R(w,s) = \tilde{R}(w,s) := w\exp(s^{-1}) + k\sqrt{\pi}\,\Phi(s^{-1/2}) - k\sqrt{\pi}. \tag{9.3.7}$$

By Theorem 9.4, the free-boundary problem (9.3.6) corresponds to the optimal stopping problem

$$u(w,s) = \sup_{\tau \leq s} \mathbb{E}\Big[R(w + W_\tau, s - \tau)\Big] \tag{9.3.8}$$

where W is a standard Brownian motion and the region $\mathcal{C}$ corresponding to the continuation region of the optimal stopping problem.

Furthermore, as $s \mapsto k\sqrt{\pi}\,\Phi(s^{-1/2}) - k\sqrt{\pi}$ is decreasing, there exists a non-negative function $w_k^*(s)$ such that

$$\mathcal{C} = \{(w,s) : w < w_k^*(s)\}.$$

In particular, we can write

$$\gamma(x_0, t_0) = x_0 + 2k\sigma t_0^{-1/2} - \sigma\sqrt{\alpha} w_k^*\Big(\frac{1}{\alpha t_0}\Big). \tag{9.3.9}$$

Hence, to justify an approximation (9.3.1), it suffices to show that as $s \to 0$,

$$w_k^*(s) \approx ks^{1/2} + \frac{1}{\sqrt{2}}s. \tag{9.3.10}$$

9.3.3 Approximation to the robust index with large information size

In this section, we will approximate $w_k^*(s)$ when s is small as stated in (9.3.10).

Recall that as $x \to \infty$,

$$\int_x^\infty \exp(-u^2/2)du \sim \frac{1}{x}\exp(-x^2/2),$$

and

$$\int_x^\infty \exp(-u^2/2)du = \sqrt{2}\Big(\frac{\sqrt{\pi}}{2} - \frac{\sqrt{\pi}}{2}\,\Phi(x/\sqrt{2})\Big)$$

Hence, $\sqrt{\pi}\,\Phi(x) - \sqrt{\pi} \sim -\frac{1}{x}\exp(-x^2)$ as $x \to \infty$. Therefore, the reward $\tilde{R}$ given in (9.3.7) can be approximated by

$$\tilde{R}(w,s) \sim \hat{R}(w,s) := (w - ks^{1/2})\exp(-s^{-1}) \qquad \text{as } s \to 0. \tag{9.3.11}$$

For the rest of this section, we will study the solution to the free-boundary problem (9.3.6) with

$$R(w,s) = \hat{R}(w,s) := (w - ks^{1/2})\exp(-s^{-1})$$

when $s \to 0$.

Remark 9.4. It is easy to see that the reward $\tilde{R}$ and $\hat{R}$ given in equations (9.3.7) and (9.3.11) satisfies condition for Theorem 9.4.

Therefore, the continuation regions in (9.3.6) corresponding to $\tilde{R}$ and $\hat{R}$ can be given by

$$\tilde{\mathcal{C}} = \big\{(w,s) : w < \tilde{w}_k^*(s)\big\} \qquad \text{and} \qquad \hat{\mathcal{C}} = \big\{(w,s) : w < \hat{w}_k^*(s)\big\}$$

As $s \to 0$, $\tilde{R}(w,s) \sim \hat{R}(w,s)$, we may approximate $\tilde{w}_k^*$ by $\hat{w}_k^*$ around $s = 0$ and use it to obtain the boundary of the optimal stopping problem (9.3.4).

We will follow Bather' comparison technique [14] to establish that for the free-boundary problem 9.3.6 with $R(w,s) = (w - ks^{1/2})\exp(s^{-1})$, the continuation region $\mathcal{C}$ can be given by $\mathcal{C} := \big\{(w,s) : w < w_k^*(s)\big\}$ where

$$w_k^*(s) \sim ks^{1/2} + \frac{1}{\sqrt{2}}s \qquad \text{as } s \to 0. \tag{9.3.12}$$

Upper bound of the free-boundary

To obtain an approximation (9.3.12), we need to establish an upper bound and a lower bound of the free-boundary of (9.3.6). We can then take the limit as $s \to 0$ to obtain the required result.

We will first establish an upper bound of $w_k^*(s)$. This requires the following comparison theorem where the proof can be found in Bather [14].

Theorem 9.8. *Let u and u' be the solution to an optimal stopping problem* (9.3.8) *with reward R and R' and continuation region $\mathcal{C}$ and $\mathcal{C}'$, respectively. Suppose that $(w_0, s_0) \notin \mathcal{C}'$ but*

(i) $R'(w_0, s_0) = R(w_0, s_0)$

(ii) $R'(w, s) \geq R(w, s)$ *for all* w *and* $s < s_0$.

Then $R(w_0, s_0) \geq u(w_0, s_0)$. In particular, $(w_0, s_0) \notin \mathcal{C}$.

Corollary 9.9. *If R' is a solution to the heat equation*

$$\frac{1}{2}\,\partial_{ww} R' - \partial_s R' = 0$$

for all $(w, s) \in \mathbb{R} \times (0, \infty)$ such that $R'(w_0, s_0) = R(w_0, s_0)$ and $R'(w, s) \geq R(w, s)$ for all w and $s < s_0$. Then $(w_0, s_0) \notin \mathcal{C}$.

Theorem 9.10. *For $R(w, s) = (w - ks^{1/2}) \exp(s^{-1})$, the boundary function $w_k^*(s)$ satisfies*

$$w_k^*(s) \leq ks^{1/2} + \frac{1}{4} s\sqrt{8 + k^2 s} + \frac{1}{4} ks^{3/2}.$$

Proof. Consider

$$R'(w, s) = \exp\Big(aw + \frac{1}{2}a^2 s + b\Big)$$

which is a solution to heat equation for all values of a and b. We will identify a and b later in terms of a fixed s_0. We will assume for now that $a > 0$.

For $w \leq ks^{1/2}$, we have $R'(w, s) > 0 \geq R(w, s)$.

For $w > ks^{1/2}$.

$$\frac{R'(w, s)}{R(w, s)} = \frac{1}{w - ks^{1/2}} \exp\Big(aw + \frac{1}{2}a^2 s + \frac{1}{s} + b\Big).$$

We define

$$K(w, s) = \frac{1}{w - ks^{1/2}} \exp\Big(aw + \frac{1}{2}a^2 s + \frac{1}{s} + b\Big).$$

It follows that

$$\partial_w K = \frac{1}{(w - ks^{1/2})^2} \exp\Big(aw + \frac{1}{2}a^2 s + \frac{1}{s} + b\Big)\Big[a(w - ks^{1/2}) - 1)\Big].$$

We can see that for a fixed s, K has a unique stationary point at

$$\hat{w}(s) = ks^{1/2} + 1/a. \tag{9.3.13}$$

Furthermore, one can check that $\partial_{ww} K(\hat{w}(s), s) > 0$. Hence, $\hat{w}(s)$ is the minimum point of K.

Define $M(s) = \min_w K(w,s)$. We have

$$M(s) = K(\hat{w}(s), s) = a\exp\Big(1 + aks^{1/2} + \frac{1}{2}a^2 s + \frac{1}{s} + b\Big),$$
$$M'(s) = \frac{a}{s^2}\exp\Big(1 + aks^{1/2} + \frac{1}{2}a^2 s + \frac{1}{s} + b\Big)\Big(\frac{1}{2}aks^{3/2} + \frac{1}{2}a^2 s^2 - 1\Big).$$

As $\frac{d}{ds}\Big(\frac{1}{2}aks^{3/2} + \frac{1}{2}a^2 s^2 - 1\Big) > 0$ for all $s > 0$, it follows from the standard algebra that M has exactly one stationary point on $(0,\infty)$. Since $M(s) \to \infty$ as $s \to 0$ and $s \to \infty$, the stationary point must be a minimum.

We fix $s_0 > 0$ and we choose a to be the positive root of the equation

$$0 = \frac{1}{2}aks_0^{3/2} + \frac{1}{2}a^2 s_0^2 - 1. \tag{9.3.14}$$

It then follows from the above argument that $s_0 = \arg\max_s M(s)$.

We then choose b such that $M(s_0) = 1$.

It now follows that for all (w,s),

$$K(w,s) \geq \min_w K(w,s) =: M(s) \geq M(s_0) = 1.$$

Therefore, for such choice of a, b, $R'(w,s) \geq R(w,s)$ for all (w,s) and $R'(w_0,s_0) = R(w_0,s_0)$.

Moreover, by (9.3.13) and (9.3.14),

$$\begin{aligned} w_0 &= ks_0^{1/2} + 1/a = ks_0^{1/2} + \frac{2\big(\frac{1}{2}s_0^2\big)}{-\frac{1}{2}ks_0^{3/2} + \sqrt{\Big(\frac{1}{2}ks_0^{3/2}\Big)^2 + 4\Big(\frac{1}{2}s_0^2\Big)}} \\ &= ks_0^2 + \frac{1}{4}ks_0^{3/2} + \frac{1}{4}s_0\sqrt{8 + k^2 s_0}. \end{aligned}$$

By Theorem 9.8, if we write $s = s_0$, we have

$$\Big(s, ks^2 + \frac{1}{4}ks^{3/2} + \frac{1}{4}s\sqrt{8 + k^2 s}\Big) \notin \mathcal{C}.$$

By Theorem 9.4, $\mathcal{C} = \big\{(w,s) : w < w_k^*(s)\big\}$. It follows that

$$w_k^*(s) \leq ks^{1/2} + \frac{1}{4}ks^{3/2} + \frac{1}{4}s\sqrt{8 + k^2 s}.$$

□

Lower bound of the free-boundary

To achieve an approximation of $w_k^*(s)$ when $s \to 0$, we also need to find an asymptotic lower bound of $w_k^*(s)$. This can be done by considering another comparison technique. The proof again can be found in Bather [14].

Theorem 9.11. *Let u be the solution to an optimal stopping problem* (9.3.8) *with reward R and continuation region $\mathcal{C}$. Let g be a continuous solution to the heat equation*

$$\frac{1}{2}\partial_{ww} g - \partial_s g = 0.$$

Define

$$A := \{(w,s) : g(w,s) > R(w,s)\}$$

Let $Z_A^{w,s}$ be a random variable taking values in $\bar{A}$ such that the path $(w+W_t, s-t)$ hits the boundary of A, ∂A where $\bar{A}$ is a closure of A and $\partial A := \bar{A}\backslash Int(A)$.

Define $u_A(w,s) = \mathbb{E}\big(R(Z_A^{w,s})\big)$. If R is continuous, then for all $(w,s) \in \bar{A}$, $u_A(w,s) = g(w,s)$. In particular, $A \subseteq \mathcal{C}$.

Theorem 9.12. *Suppose that the boundary function $w_k^*(s)$ is continuous, then the boundary function $w_k^*(s)$ satisfies*

$$\liminf_{s\downarrow 0}\Big(\frac{w_k^*(s) - ks^{1/2}}{s}\Big) \geq \frac{1}{\sqrt{2}}.$$

Proof. Consider

$$g(w,s) = c\exp(2^{1/2}b^{-1}w + b^{-2}(s-b)) - c$$

which is a solution to the heat equation for all values of b and c which will be chosen later. We assume for now that $b > 0$.

The heuristic idea of the proof proceeds as follows: for a fixed s_0 in a neighbourhood of 0, we want to find b and c depending on s_0 such that we can a function $\hat{w}(s_0)$ such that $\{(w,s_0) : w \in (0,\hat{w}(s_0))\} \subseteq \{(w,s) : g(w,s) > R(w,s)\}$ where $\hat{w}$ is a function that can be written explicitly.

It then follows from Theorem 9.11 that $\hat{w}(s_0) \leq w_k^*(s_0)$. By considering the limit as $s_0 \to 0$, we can obtain the required result.

However, in our setting $\hat{w}$ can be given explicitly in terms of s_0. Therefore, we will consider s_0 in terms of b and take $s_0 \to 0$ implicitly to establish the required result.

Define $K(w,s) = g(w,s) - R(w,s)$. We note that

$$K(w,s) = c\exp(2^{1/2}b^{-1}w + b^{-2}(s-b)) - c - (w - ks^{1/2})\exp(-\frac{1}{s}).$$

Observe that

$$K(0,s) = c\exp(b^{-2}(s-b)) - c + ks^{1/2}\exp(-\frac{1}{s}).$$

Hence, for $b, c > 0$ and $s \geq b$, $K(0,s) > 0$.

Step I: We want to find $s(b)$ and $\hat{w}(s(b))$ such that $K(\hat{w}(s(b),s(b)) = 0$ and $K(w,s(b)) > 0$ for all $w \in (0,\hat{w}(s(b)))$.

By standard algebra, one can show that $w \mapsto K(w,s)$ has a unique minimiser at

$$\hat{w}(s) = -2^{-1/2}b^{-1}(s-b) + 2^{-1/2}b\log(2^{-1/2}b/c) - 2^{-1/2}bs^{-1}.$$

It then follows that

$$M(s) := \min_w K(w,s) = K(\hat{w}(s),s) = \Big(2^{-1/2}b + ks^{1/2} - \hat{w}(s)\Big)\exp(-s^{-1}) - c.$$

Therefore,

$$M(b) = \Big(2^{-1/2}b + kb^{1/2} + 2^{-1/2} - 2^{-1/2}b\log(2^{-1/2}b/c)\Big)\exp(-b^{-1}) - c$$

For $c = 2^{-1/2}b\exp(-1-b^{-1}-k\sqrt{2}b^{-1/2}) > 0$, we have $\min_w K(w,b) = M(b) = -c < 0$.

As $s \to \infty$, one can see that $M(s) \to \infty$. As M is a smooth and continuous function on $[b,\infty)$, by the intermediate value theorem, there exists a minimal $s(b)$ such that $M(s(b)) = 0$ and $M(s) < 0$ for $s \in [b, s(b))$.

For $c = 2^{-1/2}b\exp(-1-b^{-1}-k\sqrt{2}b^{-1/2})$, we can write

$$\hat{w}(s) = 2^{-1/2}b\Big(1 - s(b^{-1}-s^{-1})^2\Big) + kb^{1/2}$$

and

$$M(s) = \Big(2^{-1/2}bs(b^{-1}-s^{-1})^2 + k(s^{1/2}-b^{1/2})\Big)\exp(-s^{-1}) - 2^{-1/2}b\exp(-1-b^{-1}-k2^{1/2}b^{-1/2}). \tag{9.3.15}$$

Therefore, $s(b)$ is the unique solution to

$$2^{-1/2}bs(b^{-1}-s^{-1})^2 = 2^{-1/2}b\exp(-1+s^{-1}-b^{-1}-2^{1/2}kb^{-1/2}) - k(s^{1/2}-b^{1/2}).$$

By substituting the expression above into $\hat{w}(s)$, we have

$$\hat{w}(s(b)) = 2^{-1/2}b\Big(1 - \exp(-1+s(b)^{-1}-b^{-1}-\sqrt{2}kb^{-1/2})\Big) + ks(b)^{1/2}.$$

As $s(b) > b$, $s(b)^{-1} - b^{-1} < 0$. Therefore, $\hat{w}(s(b)) > 0$.

As $K(0,s(b)) > 0$ and $K(\hat{w}(s(b)),s(b)) = 0$, by a strict convexity of $w \mapsto K(w,s)$, it follows that $K(w,s(b)) > 0$ for all $w \in (0,\hat{w}(s(b)))$.

Therefore, for any fixed b, it follows from Theorem 9.11 that

$$\begin{aligned}\Big\{(w,s(b)) : w \in \big(0,\hat{w}(s(b))\big)\Big\} &\subseteq \Big\{(w,s) : K(w,s) > 0\Big\} \\ &= \Big\{(w,s) : g(w,s) > R(w,s)\Big\} \subseteq \mathcal{C}.\end{aligned}$$

In particular, we have

$$\hat{w}(s(b)) \le w_k^*(s(b)). \tag{9.3.16}$$

Step II: For a fixed $s_0 > 0$ in the neighbourhood of 0, we want to choose b such that $s(b) = s_0$. We can then consider the limit as $s_0 \to 0$ to establish the required result.

Fix $s_0 > 0$. We want to choose $b = b(s_0)$ such that $M(s_0) = 0$ where M is given as a function of b and s in (9.3.15). (i.e. we want to find b such that $s(b) = s_0$.) Hence, b must satisfies

$$\Big(2^{-1/2}bs_0(b^{-1}-s_0^{-1})^2 + k(s_0^{1/2}-b^{1/2})\Big) = 2^{-1/2}b\exp(-1+s_0^{-1}-b^{-1}-\sqrt{2}kb^{-1/2}). \tag{9.3.17}$$

For $b = s_0$, $LHS = 0$ and $RHS > 0$. For $b \to 0$, $LHS \to \infty$ and $RHS \to 0$. Hence, for a fixed $s_0 > 0$, there exists $b = b(s_0) \in (0, s_0)$ such that for such value of b, $M(s_0) = 0$.

We recall that for such $b = b(s_0)$,

$$\hat{w}(s_0) = 2^{-1/2} b\Big(1 - \exp(-1 + s_0^{-1} - b^{-1} - \sqrt{2}kb^{-1/2})\Big) + ks_0^{1/2}.$$

Hence, to prove the required result, it suffices to show that

$$\begin{aligned}\frac{1}{\sqrt{2}} &= \liminf_{s\downarrow 0}\Big(\frac{w_k^*(s_0) - ks_0^{1/2}}{s_0}\Big)\\ &= \liminf_{s_0\downarrow 0}\left(\frac{2^{-1/2}b\Big(1 - \exp(-1 + s_0^{-1} - b^{-1} - \sqrt{2}kb^{-1/2})\Big)}{s_0}\right).\end{aligned}$$

As $s_0 \to 0$, $\big|2^{-1/2}bs_0(b^{-1} - s_0^{-1})^2\big|/s_0^{1/2} \to \infty$ but $k(s_0^{1/2} - b^{1/2}) = \mathcal{O}(s_0^{1/2})$. Hence,

$$2^{-1/2}bs_0(b^{-1} - s_0^{-1})^2 + k(s_0^{1/2} - b^{1/2}) \sim 2^{-1/2}bs_0(b^{-1} - s_0^{-1})^2 \qquad \text{as } s_0 \to 0.$$

Moreover, $b^{-1} + \sqrt{2}kb^{-1/2} \sim b^{-1}$ as $b \to 0$. Hence,

$$2^{-1/2}b\exp(-1 + s_0^{-1} - b^{-1} - \sqrt{2}kb^{-1/2}) \sim 2^{-1/2}b\exp(-1 + s_0^{-1} - b^{-1}).$$

Therefore, by (9.3.17), we have

$$2^{-1/2}bs_0(b^{-1} - s_0^{-1})^2 \sim 2^{-1/2}b\exp(-1 + s_0^{-1} - b^{-1})$$

Define $\xi = b^{-1} - s_0^{-1}$. It follows from above that $s_0 \sim \xi^{-2}\exp(-1 - \xi)$. Therefore, $b = \Big(\xi + s_0^{-1}\Big)^{-1} \sim \Big(\xi + \xi^2\exp(1 + \xi)\Big)^{-1}$. As $s_0 \to 0$ and $s_0 \sim \xi^{-2}\exp(-1 - \xi)$, we must have $\xi \to \infty$ as $s_0 \to 0$. It then follows that

$$\begin{aligned}\hat{w}(s_0) &= 2^{-1/2}bs_0^{-1}\Big(1 - \exp(-1 + s_0^{-1} - b^{-1} - \sqrt{2}kb^{-1/2})\Big)\\ &\sim\ 2^{-1/2}bs_0^{-1}\Big(1 - \exp(-1 + s_0^{-1} - b^{-1})\Big)\\ &\sim\ 2^{-1/2}\frac{\xi^2\exp((1 + \xi))}{\xi + \xi^2\exp(1 + \xi)}\Big(1 - \exp(-1 - \xi)\Big) \to 1/\sqrt{2} \qquad \text{as } \xi \to \infty.\end{aligned}$$

Therefore, by inequality (9.3.16),

$$\frac{1}{\sqrt{2}} = \liminf_{s_0\downarrow 0}\frac{\hat{w}(s_0) - ks_0^{1/2}}{s_0} \le \liminf_{s_0\downarrow 0}\frac{w_k^*(s_0) - ks_0^{1/2}}{s_0}.$$

□

Corollary 9.13. *For $R(w, s) = (w - ks^{1/2})\exp(s^{-1})$, the boundary function $w_k^*(s)$ of (9.2.5) satisfies*

$$\lim_{s\downarrow 0}\frac{w_k^*(s) - ks^{1/2}}{s} = \frac{1}{\sqrt{2}}.$$

Proof. The result follows from Theorem 9.10 and Theorem 9.12. □

Appendix A

Interim results for Asymptotic Randomised Control

This appendix presents interim results for the Asymptotic Randomised Control in Chapter 3.

A.1 Interim results

Lemma A.1. *(Grönwall's inequality) Let* (y_t) *and* (c_t) *be non-negative sequences and* $\rho > 0$*. Suppose that*

$$y_T \leq c_T + \rho \sum_{t=1}^{T-1} y_t \qquad \text{for all} \quad T \geq 1.$$

Then

$$y_T \leq c_T + \rho \sum_{t=1}^{T-1} c_t (1+\rho)^{T-t-1}.$$

Corollary A.2. *Let* (b_t) *be a non-negative sequence and* $c, \rho > 0$ *be such that*

$$b_T \leq c + \rho \sum_{t=1}^{T-1} \beta^{T-t} b_t.$$

Then

$$b_T \leq c\Big(1 + \Big(\frac{\rho}{1+\rho}\Big) \sum_{t=1}^{T-1} \big(\beta(1+\rho)\big)^t\Big). \tag{A.1.1}$$

In particular, if $0 < \rho = (1-\beta) < (1-\beta)/\beta$*, then*

$$b_T \leq \frac{c(1-\beta)}{1-\beta(1+\rho)} = \frac{c}{(1-\beta)}.$$

Proof. Define $y_t := \beta^{-t} b_t$ and $c_t = c\beta^{-t}$. Then $y_T \le c_T + \rho \sum_{t=1}^{T-1} y_t$. Hence, by Grönwall's inequality, $y_T \le c_T + \rho \sum_{t=1}^{T-1} c_t (1+\rho)^{T-t-1}$. In particular,

$$\begin{aligned} b_T &\le c + c\rho \sum_{t=1}^{T-1} \beta^{T-t} (1+\rho)^{T-t-1} = c\Big(1 + \rho\beta \sum_{t=1}^{T-1} \beta^{T-t-1} (1+\rho)^{T-t-1}\Big) \\ &= c\Big(1 + \rho\beta \sum_{t=0}^{T-2} \big(\beta(1+\rho)\big)^t\Big). \end{aligned}$$

Finally, for $0 < \rho < (1-\beta)/\beta$, $0 < \beta(1+\rho) < \beta(1 + (1-\beta)/\beta) = 1$. Hence,

$$b_T = c\Big(1 + \rho\beta \sum_{t=0}^{T-2} \big(\beta(1+\rho)\big)^t\Big) \le c\Big(1 + \frac{\rho\beta}{1-\beta(1+\rho)}\Big) = \frac{c(1-\beta)}{1-\beta(1+\rho)}.$$

□

Lemma A.3. *There exist $P_2 \in \mathcal{P}_2^-$ and $q_2 \in \mathcal{P}_2^+$ such that, for any $m, \Delta m \in \mathbb{R}^G, d, \Delta d \in \mathbb{R}^H$ with $d \in \mathcal{D}(\lambda)$ and $|\Delta d| \le C|d|^2$,*

$$\begin{aligned} \Big| \mathrm{S}_{\max}\big(\alpha_T(m+\Delta m, d+\Delta d)\big) - \mathrm{S}_{\max}\big(f(m+\Delta m, d+\Delta d) + F_T(m,d)\big)\Big| \\ \le (1-\beta)^{-2} P_2(\lambda) q_2(|d|)\big(|\Delta m| + |d|\big). \end{aligned}$$

Proof. Recall that $\mathrm{S}_{\max}(a) = \sup_u \Big\{ \sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big\}$. Hence,

$$\begin{aligned} |\mathrm{S}_{\max}(a+\Delta a) - \mathrm{S}_{\max}(a)| &= \Big| \sup_u \Big\{ \sum_{i=1}^K u_i (a_i + \Delta a_i) + \lambda \mathcal{H}(u) \Big\} - \sup_u \Big\{ \sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big\} \Big| \\ &\le \sup_u \Big| \sum_{i=1}^K u_i (a_i + \Delta a_i) + \lambda \mathcal{H}(u) - \sum_{i=1}^K u_i a_i + \lambda \mathcal{H}(u) \Big| = \sup_u \Big| \sum_{i=1}^K u_i \Delta a_i \Big| \le |\Delta a|. \end{aligned}$$

By considering $\Delta a = \Delta F_T$, the result follows from Corollary 3.7. □

Lemma A.4. *There exists $P_2 \in \mathcal{P}_2^-$ such that, for any function $a_T(m,d) = f(m,d) + \phi_T$, where ϕ_T is a constant,*

$$\begin{aligned} &\big(\mathrm{S}_{\max} \circ a_T\big)(m+\Delta m, d+\Delta d) - \big(\mathrm{S}_{\max} \circ a_T\big)(m,d) \\ &\quad = \Big\langle \partial_d \big(\mathrm{S}_{\max} \circ a_T\big), \Delta d \Big\rangle + \frac{1}{2} \Big\langle \partial_m^2 \big(\mathrm{S}_{\max} \circ a_T\big), \Delta m \Delta m^\top \Big\rangle \\ &\qquad + \Big\langle \partial_m \big(\mathrm{S}_{\max} \circ a_T\big), \Delta m \Big\rangle + \Delta_1 S_T(m,d), \end{aligned}$$

where

$$|\Delta_1 S_T(m,d)| \le P_2(\lambda)(|\Delta m|^3 + |d|^3)$$

for all $m, \Delta m \in \mathbb{R}^G, d, \Delta d \in \mathbb{R}^H$ with $|\Delta d| \le C|d|^2$.

Proof. Recall that for any $x, y > 0$, we have $x^k y^l \le (x^{k+l} + y^{k+l})$ for $k, l \ge 1$. Hence, by Taylor's theorem and a similar argument to Lemma 3.5, it suffices to show that

$\partial_d^2\big(\mathrm{S}_{\max}\circ a_T\big), \partial_d\,\partial_m\big(\mathrm{S}_{\max}\circ a_T\big)$ and the third derivatives of $\big(\mathrm{S}_{\max}\circ a_T\big)$ are uniformly bounded by a quadratic in $1/\lambda$.

By standard algebra, one can bound $\partial_d^2\big(\mathrm{S}_{\max}\circ a_T\big)$, $\partial_d\,\partial_m\big(\mathrm{S}_{\max}\circ a_T\big)$ and the third derivatives of $\big(\mathrm{S}_{\max}\circ a_T\big)$ by $\sum_{j=1}^{k}|\mathcal{C}_j f(m,d)|\big|\mathcal{D}_j\,\mathrm{S}_{\max}\big(a_T(m,d)\big)\big|$ where $\mathcal{C}_j$ and $\mathcal{D}_j$ are differential operators up to the third order with respect to (m,d) and a, respectively.

By Assumption 3.3, $|\mathcal{C}_j f(m,d)|$ is uniformly bounded. By a similar argument to the proof of Lemma 3.5, the derivatives of $\mathrm{S}_{\max}$ with respect to a up to the third order can be bounded by a quadratic polynomial of $1/\lambda$.

By combining those bounds, the result follows. □

Lemma A.5. *We have* $\sup_{m,d}|V_T(m,d)-V_\infty(m,d)|\to 0$ *as* $T\to\infty$.

Proof. Let B be a uniform bound on $\mathcal{H}(u)$. By Assumption 3.3, we have

$$\begin{aligned}&|V_T(m,d)-V_\infty(m,d)|\\&\quad\le\sup_{U\in\mathcal{U}}\left|\mathbb{E}_{m,d}\Big[\sum_{t=T}^{\infty}\beta^t\Big(f\big(M_t^U,D_t^U,A_{t+1}\big)+\lambda\mathcal{H}(U_{t+1})\Big)\Big]\right|\le\Big(\frac{\beta^T}{1-\beta}\Big)(C+\lambda B).\end{aligned}$$

Hence, $V_T\to V_\infty$ uniformly as $T\to\infty$. □

Proposition A.6 (Tauberian theorem)**.** *Suppose that* g_t *is a sequence of uniformly bounded functions such that* $g_t\to g$ *uniformly. Then*

$$\sum_{t=1}^{T}\beta^{T-t}g_t\to\Big(\frac{1}{1-\beta}\Big)g\qquad\textit{uniformly}\qquad\textit{as}\qquad T\to\infty.$$

Proof. Fix $\epsilon>0$. We can find $s>0$ such that for all $t\ge s$, $\|g_t-g\|_\infty\le\epsilon$.

Since $g_t\to g$ and (g_t) is uniformly bounded, the sequence $f_n:=g_n-g$ is uniformly bounded. Hence, there exists $T_0>s$ such that for all $t>T_0-s$, $\beta^{t/2}\|g_n-g\|_\infty\le\epsilon$ for all $n\in\mathbb{N}$.

Therefore, for $T\ge T_0$,

$$\begin{aligned}\Big\|\sum_{t=1}^{T}\beta^{T-t}g_t-\Big(\frac{1}{1-\beta}\Big)g\Big\|_\infty&=\Big\|\sum_{t=1}^{T}\beta^{T-t}(g_t-g)+\Big(\frac{\beta^T}{1-\beta}\Big)g\Big\|_\infty\\&\le\sum_{t=1}^{s}\beta^{T-t}\big\|g_t-g\big\|_\infty+\sum_{t=s+1}^{T}\beta^{T-t}\big\|g_t-g\big\|_\infty+\Big(\frac{\beta^T}{1-\beta}\Big)\|g\|_\infty.\end{aligned}$$

Observe that

$$\sum_{t=1}^{s}\beta^{T-t}\big\|g_t-g\big\|_\infty=\sum_{t=T-s}^{T-1}\beta^{t/2}\big(\beta^{t/2}\big\|g_{T-t}-g\big\|_\infty\big)\le\epsilon\sum_{t=T-s}^{T-1}\beta^{t/2}\le\Big(\frac{1}{1-\beta^{1/2}}\Big)\epsilon$$

and

$$\sum_{t=s+1}^{T}\beta^{T-t}\big\|g_t-g\big\|_\infty\le\sum_{t=s+1}^{T}\beta^{T-t}\epsilon\le\Big(\frac{1}{1-\beta}\Big)\epsilon.$$

Hence, it follows that

$$\limsup_{T\to\infty}\Big\|\sum_{t=1}^{T}\beta^{T-t}g_t-\Big(\frac{1}{1-\beta}\Big)g\Big\|_\infty\le\Big(\frac{1}{1-\beta^{1/2}}\Big)\epsilon+\Big(\frac{1}{1-\beta}\Big)\epsilon.$$

As ϵ is arbitrary, the result follows. □

Corollary A.7. *Let α be the function defined in* (3.4.12)*. As $T\to\infty$,*

$$\sup_{|d|\le h,m}\Big|\sum_{t=1}^{T}\beta^{T-t}\big(\mathrm{S}_{\max}\circ\alpha_t\big)(m,d)-\Big(\frac{1}{1-\beta}\Big)\big(\mathrm{S}_{\max}\circ\alpha\big)(m,d)\Big|\to 0.$$

Proof. By Theorem 3.12, $\alpha_t\to\alpha$ uniformly on $\{(m,d)\in\mathbb{R}^{G+H}:|d|\le h\}$.

Since the derivative of $\mathrm{S}_{\max}(a)$ is $\nu(a)$, which is uniformly bounded, $a\mapsto\mathrm{S}_{\max}(a)$ is a Lipschitz function. Hence, by the mean value inequality, $\big(\mathrm{S}_{\max}\circ\alpha_t\big)\to\big(\mathrm{S}_{\max}\circ\alpha\big)$ uniformly on $\{(m,d)\in\mathbb{R}^{G+H}:|d|\le h\}$. The result follows from Proposition A.6. □

A.2 Proofs of stated results

Proof of Theorem 3.2. $\underline{(i)\Rightarrow(ii)}$: Fix i. Consider $a=(N+\epsilon)e^{(i)}+\sum_{j\neq i}r_je^{(j)}$ where $r_j\in\mathbb{R}$ for all $j\neq i$. By (i), $S(a)+N\ge\max(N+\epsilon,r_j)\ge N+\epsilon$. Hence, $S(a)\ge\epsilon>0$.

As ϵ is arbitrary, it follows that $\mathbb{R}^{i-1}\times(N,\infty)\times\mathbb{R}^{K-i}\subseteq\mathcal{A}_S^c$. The result then follows by considering intersection over all i.

$\underline{(ii)\Rightarrow(iii)}$: By Theorem 3.1, we can write $S(a)=\sup_{u\in\Delta^K}\Big(\sum_{i=1}^K u_ia_i+\mathcal{H}_{max}(u)\Big)$, where $\mathcal{H}_{max}(u):=-\sup_{a\in\mathcal{A}_S}\Big(\sum_{i=1}^K u_ia_i\Big)$ and $\mathcal{A}_S:=\{a\in\mathbb{R}^K:S(a)\le 0\}$.

As $\mathcal{A}_S\subseteq(-\infty,N]^K$ and $u\in\Delta^K$,

$$\mathcal{H}_{max}(u)\ge-\sup_{a\in(-\infty,N]^K}\Big(\sum_{i=1}^K u_ia_i\Big)\ge -N.$$

Moreover, by (3.4.5), we have $\sup_{u\in\Delta^K}\mathcal{H}_{max}(u)\le S(0)$. Therefore, $\mathcal{H}_{max}$ is bounded.

$\underline{(iii)\Rightarrow(iv)}$ Fix $a\in\mathbb{R}^K$ and define $i^*\in\arg\max_i a_i$. Then

$$\begin{aligned}-\lambda\sup_{u\in\Delta^K}|\mathcal{H}(u)|\le\lambda\mathcal{H}(e^{(i^*)})&=\sum_{i=1}^K(e^{(i^*)})_ia_i+\lambda\mathcal{H}(e^{(i^*)})-\max_i a_i\\&\le\mathrm{S}^\lambda_{\max}(a)-\max_i a_i\le\sup_{u\in\Delta^K}\Big(\sum_{i=1}^K u_ia_i\Big)+\lambda\sup_{u\in\Delta^K}|\mathcal{H}(u)|-\max_i a_i=\lambda\sup_{u\in\Delta^K}|\mathcal{H}(u)|.\end{aligned}$$

Hence, $\sup_{a\in\mathbb{R}}|\mathrm{S}^\lambda_{\max}(a)-\max_i a_i|\le\lambda\sup_{u\in\Delta^K}|\mathcal{H}(u)|\to 0$ as $\lambda\downarrow 0$.

$\underline{(iv)\Rightarrow(i)}$ Find $N>0$ such that

$$1\ge\sup_{a\in\mathbb{R}}|S^{1/N}_{\max}(a)-\max_i a_i|=\frac{1}{N}\sup_{a\in\mathbb{R}}|S(Na)-\max_i Na_i|=\frac{1}{N}\sup_{a\in\mathbb{R}}|S(a)-\max_i a_i|.$$

By rearranging the inequality above, the result follows. □

Proof of Lemma 3.5. As $f^{(i)}$ is 3-times differentiable with bounded derivatives, it follows from Taylor's approximation (or the mean value inequality) that there exists $\tilde{P}_2 \in \mathcal{P}_2^-$, a quadratic polynomial in $1/\lambda$, such that

$$\begin{aligned}\mathcal{B}(a+\Delta a, m+\Delta m, d+\Delta d) &= \mathcal{B}(a,m,d) + \Delta\mathcal{B},\\ \mathcal{M}(a+\Delta a, m+\Delta m, d+\Delta d) &= \mathcal{M}(a,m,d) + \Delta\mathcal{M},\\ \Sigma(a+\Delta a, m+\Delta m, d+\Delta d) &= \Sigma(a,m,d) + \Delta\Sigma,\end{aligned}$$

where $|\Delta\mathcal{B}|, |\Delta\mathcal{M}|, |\Delta\Sigma| \le \tilde{P}_2(\lambda)\big(|\Delta a| + |\Delta m| + |\Delta d|\big)$.

The polynomial is of order $1/\lambda^2$ because we need to consider the first derivatives of $\mathcal{B}$, $\mathcal{M}$ and Σ. These terms can be written as a linear function of $\partial_a \mathrm{S}_{\max}(a)$ and $\partial_a^2 \mathrm{S}_{\max}(a)$, so taking the first order derivative for these terms w.r.t. (a,m,d), results in a linear function involving $\partial_a^3 \mathrm{S}_{\max}(a)$. As $\mathrm{S}_{\max}(a) = \lambda S(a/\lambda)$ where S is a convex conjugate of $-\mathcal{H}$, it follows that $\partial_a^3 \mathrm{S}_{\max}(a) = 1/\lambda^2\big(\partial_y^3 S(y)|_{y=a/\lambda}\big)$. In particular, as S is C_b^3, we can bound the first derivatives of $\mathcal{B}$, $\mathcal{M}$ and Σ by a quadratic polynomial of $1/\lambda$.

Furthermore, we can also find $\tilde{P}_1(\lambda)$, an affine function in $1/\lambda$, such that

$$|\mathcal{B}|, |\mathcal{M}|, |\Sigma| \le \tilde{P}_1(\lambda).$$

By similar arguments as above, we have

$$b^{(i)}(m+\Delta m, d+\Delta d) = b^{(i)}(m,d) + \Delta b,$$

where

$$\begin{aligned}|\Delta b| &\le \sup_{\tilde{d}\in[d,d+\Delta d],m}\Big(|\,\partial_d b^{(i)}(m,\tilde{d})|\Big)|\Delta d| + \Big(\sup_{\tilde{d}\in[d,d+\Delta d],m}|\,\partial_m b^{(i)}(m,d)|\Big)|\Delta m|\\ &\le C|d+\Delta d||\Delta d| + C|d+\Delta d|^2|\Delta m| \le \tilde{C}_1 q_2(|d|)\big(|\Delta m| + |d|\big).\end{aligned}$$

for some constant $\tilde{C}_1$ and $q_2 \in \mathcal{P}_2^-$. The last inequality follows from triangle inequality and the fact the $|\Delta d| \le C|d|^2$.

Similarly, from Assumption 3.3, there exist constants $\tilde{C}_2, \tilde{C}_3$ and polynomials $\hat{q}_2, \tilde{q}_2 \in \mathcal{P}_2^+$ such that

$$\begin{aligned}\mu^{(i)}(m+\Delta m, d+\Delta d) &= \mu^{(i)}(m,d) + \Delta\mu,\\ \sigma^{(i)}(m+\Delta m, d+\Delta d)\sigma^{(i)}(m+\Delta m, d+\Delta d)^\top &= \sigma^{(i)}(m,d)\sigma^{(i)}(m,d)^\top + \Delta\sigma\sigma^\top\end{aligned}$$

where

$$|\Delta\mu| \le \tilde{C}_2\hat{q}_2(|d|)\big(|\Delta m| + |d|\big) \qquad \text{and} \qquad |\Delta\sigma\sigma^\top| \le \tilde{C}_3\tilde{q}_2(|d|)\big(|\Delta m| + |d|\big).$$

Moreover, we also recall that

$$|b^{(i)}(d)|, |\mu^{(i)}(d)|, |\sigma^{(i)}(d)\sigma^{(i)}(d)^\top| \le C|d|^2.$$

By linearity of the inner product we obtain

$$\begin{aligned}\Delta L_i &:= L_i(a+\Delta a, m+\Delta m, d+\Delta d) - L_i(a,m,d)\\ &= \langle \Delta\mathcal{B}; b^{(i)}(m,d)\rangle + \langle \mathcal{B}; \Delta b\rangle + \langle \Delta\mathcal{B}; \Delta b\rangle + \langle \Delta\mathcal{M}; \mu^{(i)}(m,d)\rangle + \langle \mathcal{M}; \Delta\mu\rangle + \langle \Delta\mathcal{M}; \Delta\mu\rangle\\ &\quad + \langle \Delta\Sigma; \sigma^{(i)}(m,d)\sigma^{(i)}(m,d)^\top\rangle + \langle \Sigma; \Delta\sigma\sigma^\top\rangle + \langle \Delta\Sigma; \Delta\sigma\sigma^\top\rangle.\end{aligned}$$

Hence,

$$\begin{aligned}|\Delta L_i| &= |\Delta\mathcal{B}||b^{(i)}(m,d)| + |\mathcal{B}||\Delta b| + |\Delta\mathcal{B}||\Delta b| + |\Delta\mathcal{M}||\mu^{(i)}(m,d)| + |\mathcal{M}||\Delta\mu| + |\Delta\mathcal{M}||\Delta\mu|\\ &\quad + |\Delta\Sigma||\sigma^{(i)}(m,d)\sigma^{(i)}(m,d)^\top| + |\Sigma||\Delta\sigma\sigma^\top| + |\Delta\Sigma||\Delta\sigma\sigma^\top|.\end{aligned}$$

By substituting all inequalities, the result follows. □

Proof of Lemma 3.6. Write $a_t := \alpha_t(m,d)$ and $\Delta a_t = \alpha_t(m+\Delta m, d+\Delta d) - \alpha_t(m,d)$. We define

$$R_t := |\Delta a_t| \qquad \text{and} \qquad S_t := |\Delta l_t| = \big|L(a_t+\Delta a_t, m+\Delta m, d+\Delta d) - L(a_t,m,d)\big|.$$

By Assumptions 3.3, together with Taylor's approximation, we can write

$$f(m+\Delta m, d+\Delta d) = f(m,d) + \Delta f,$$

where $|\Delta f| \le C(|\Delta m| + |\Delta d|) \le C(|\Delta m| + C|d|^2)$ for some constant $C > 0$.

By definition of α_T, we can write

$$\Delta a_T = \Delta f + \sum_{t=1}^{T-1} \beta^{T-t}\Delta l_t.$$

Therefore,

$$R_T \le C(|\Delta m| + C|d|^2) + \sum_{t=1}^{T-1} \beta^{T-t} S_t.$$

By Lemma 3.5, there exist $P_2(\lambda)$ and $q_2(h)$ such that R_T and S_T satisfy

$$S_T \le P_2(\lambda) q_2(|d|)\big(R_T + |\Delta m| + |d|\big).$$

For $t \ge 0$, define $Q_t = R_t + |\Delta m| + |d|$. As $d \in \mathcal{D}(\lambda)$, we know $P_2(\lambda)q_2(|d|) \le \rho$ where $\rho = 1-\beta$ and it follows that $S_t \le \rho Q_t$ for all $t \ge 0$. Hence,

$$Q_T \le \Big(C(|\Delta m| + C|d|^2) + |\Delta m| + |d|\Big) + \rho \sum_{t=1}^{T-1} \beta^{T-t} Q_t.$$

By Corollary A.2,

$$Q_T \le \Big(C(|\Delta m| + C|d|^2) + |\Delta m| + |d|\Big)(1-\beta)^{-1}.$$

Therefore,

$$S_T \le P_2(\lambda) q_2(|d|)\Big(C(|\Delta m| + C|d|^2) + |\Delta m| + |d|\Big)(1-\beta)^{-1}.$$

Finally, we can combine terms to obtain the required result. □

Proof of Corollary 3.9. By Assumption 3.2, $|\Delta^{(i)}d| = |b^{(i)}(d)| \le C|d|^2$. By definition of $\Delta^{(i)}m$ we have

$$\begin{aligned}\mathbb{E}[\Delta^{(i)}m] &= \mu^{(i)}(m,d),\\ \mathbb{E}[\Delta^{(i)}m\Delta^{(i)}m^\top] &= \mu^{(i)}(m,d)\mu^{(i)}(m,d)^\top + \sigma^{(i)}(m,d)\sigma^{(i)}(m,d)^\top.\end{aligned}$$

By Assumption 3.2 again, we have

$$\begin{aligned}\left|\mu^{(i)}(m,d)\mu^{(i)}(m,d)^\top\right| &\le \left|\mu^{(i)}(m,d)\right|^2 \le C^2|d|^4,\\ \mathbb{E}[|\Delta^{(i)}m|^3] &\le |\sigma^{(i)}(d)|^3\mathbb{E}|Z|^3 \le C^3\mathbb{E}|Z|^3|d|^3.\end{aligned}$$

By Lemma 3.3 and 3.8, we can rearrange the above (in)equalities to obtain the required result. □

Proof of Theorem 3.14. Fix $\lambda > 0$ and $d \in \mathcal{D}(\lambda)$.

Let $B < \infty$ be such that $|\mathcal{H}| \le B$. One can easily see from (3.4.2) and (3.5.1)) that

$$V_\infty^\lambda(m,d) + \lambda\Big(\frac{1}{1-\beta}\Big)B \ge V(m,d) \ge V_\infty^\lambda(m,d) - \lambda\Big(\frac{1}{1-\beta}\Big)B.$$

Hence,

$$Q^\lambda(u,m,d) + \lambda\Big(\frac{1}{1-\beta}\Big)B \ge Q(u,m,d) \ge Q^\lambda(u,m,d) - \lambda\Big(\frac{1}{1-\beta}\Big)B,$$

where $Q^\lambda(u,m,d) \coloneqq \sum_{i=1}^K \left(u_i\Big(f(m,d,i) + \beta\mathbb{E}[V_\infty^\lambda(\Phi(m,d,i,Z))]\Big) + \lambda\mathcal{H}(u)\right)$.

Fix $\epsilon > 0$. By Lemma A.5, we can find $\tilde{T}$ such that for all $T \ge \tilde{T}$, we have $\sup_{m,d}|V_T^\lambda(m,d) - V_\infty^\lambda(m,d)| \le \epsilon$.

We define $Q_T^\lambda(u,m,d) \coloneqq \sum_{i=1}^K \Big(u_i\Big(f(m,d,i) + \beta\mathbb{E}[V_{T-1}^\lambda(\Phi(m,d,i,Z))]\Big) + \lambda\mathcal{H}(u)\Big)$. It follows that for $T \ge \tilde{T}+1$, $\left|Q^\lambda(u,m,d) - Q_T^\lambda(u,m,d)\right| \le \beta\epsilon$.

We now define

$$\hat{Q}_T^\lambda(u,m,d) \coloneqq \sum_{i=1}^K \left(u_i\Big(\alpha_T^{\lambda,(i)}(m,d) + \sum_{t=1}^{T-1}\beta^{T-t}\big(\mathrm{S}_{\max}^\lambda \circ \alpha_t^\lambda\big)(m,d)\Big) + \lambda\mathcal{H}(u)\right).$$

As $t \mapsto |D_t^U|$ is non-increasing, by a similar argument as in the proof of Theorem 3.10 to approximate $\mathbb{E}[V_{T-1}^\lambda(\Phi(m,d,i,Z))]$, we can show that there exists a polynomial $\tilde{P}_2 \in \mathcal{P}_2^-$ and $q_3 \in \mathcal{P}_3^+$ such that $\left|\hat{Q}_T^\lambda(u,m,d) - Q_T^\lambda(u,m,d)\right| \le (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|)$.

Hence, it follows that for all $u \in \Delta^K$ and $m,d \in \mathbb{R}^d$,

$$\left|\hat{Q}_T^\lambda(u,m,d) - Q(u,m,d)\right| \le \lambda(1-\beta)^{-1}B + (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|) + \beta\epsilon.$$

Therefore, by the Dynamic Programming Principle, for $T \geq \tilde{T} + 1$,

$$\begin{aligned} V(m,d) &= \sup_{u \in \Delta^K} Q(u,m,d) \\ &\leq \sup_{u \in \Delta^K} \hat{Q}_T^\lambda(u,m,d) + \Big(\lambda(1-\beta)^{-1}B + (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|) + \beta\epsilon\Big) \\ &= \hat{Q}_T^\lambda\Big(\nu^\lambda\big(\alpha_T^\lambda(m,d)\big),m,d\Big) + \Big(\lambda(1-\beta)^{-1}B + (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|) + \beta\epsilon\Big) \\ &\leq Q\Big(\nu^\lambda\big(\alpha_T^\lambda(m,d)\big),m,d\Big) + 2\Big(\lambda(1-\beta)^{-1}B + (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|) + \beta\epsilon\Big). \end{aligned}$$

As $u \mapsto Q(u,m,d)$ and $a \mapsto \nu^\lambda(a)$ are continuous and $\alpha_T^\lambda(m,d) \to \alpha^\lambda(m,d)$, we can take $T \to \infty$ and then take $\epsilon \to 0$ to obtain

$$V(m,d) \leq Q\Big(\nu^\lambda\big(\alpha^\lambda(m,d)\big),m,d\Big) + 2\Big(\lambda(1-\beta)^{-1}B + (1-\beta)^{-4}\tilde{P}_2(\lambda)q_3(|d|)\Big).$$

Now, suppose that $\tilde{P}_2(\lambda) = \tilde{b}(1+\lambda^{-2})$ and $P_2(\lambda) = b(1+\lambda^{-2})$ where P_2 is the bound in Definition 3.7. Hence, there exists $C > 0$ such that

$$V(m,d) \leq Q\Big(\nu^\lambda\big(\alpha^\lambda(m,d)\big),m,d\Big) + 2\Big(\lambda(1-\beta)^{-1}B + C(1-\beta)^{-4}P_2(\lambda)q_3(|d|)\Big).$$

As $\lambda(m,\tilde{d})$ is consistent, for any $\tilde{d}$ such that $|\tilde{d}| \leq h^*$, we have $P_2(\lambda(m,\tilde{d}))q_2(|\tilde{d}|) < (1-\beta)$. Therefore, $(1-\beta)^{-4}P_2(\lambda(m,\tilde{d}))q_3(|\tilde{d}|) < (1-\beta)^{-3}q_3(|\tilde{d}|)/q_2(|\tilde{d}|)$.

For any $q_2 \in \mathcal{P}_2^+$ and $q_3 \in \mathcal{P}_3^+$, we can find $q_1 \in \mathcal{P}_1^+$ such that $q_3(x)/q_2(x) \leq q_1(x)$ for all $x > 0$. In particular, as $\tilde{d} \in \mathcal{D}\big(\lambda(m,\tilde{d})\big)$, we can find $q_1 \in \mathcal{P}_1^+$ such that

$$V(m,\tilde{d}) \leq Q\Big(\nu^\lambda\big(\alpha^\lambda(m,\tilde{d})\big),m,\tilde{d}\Big) + 2\lambda(m,\tilde{d})\Big(\frac{1}{1-\beta}\Big)B + 2C(1-\beta)^{-3}q_1(|\tilde{d}|).$$

Hence, we obtain a constant which depends only on β, as required. □

Proof of Corollary 3.15. We define a strategy $\tilde{U}^{\lambda,n}$ which agrees with the strategy U^λ up to time n and follows an optimal control afterwards.

As f is bounded, one can show that $\sup_{m,d}\Big|V^{U^\lambda}(m,d) - V^{\tilde{U}^{\lambda,n}}(m,d)\Big| \to 0$ as $n \to \infty$.

By Theorem 3.14 and the Dynamic Programming principle, we can show by induction that

$$V(m,d) \leq V^{\tilde{U}^{\lambda,n}}(m,d) + C\mathbb{E}\Big(\sum_{t=0}^{n-1}\beta^t\Big(\sup_m \lambda(m, D_t^{m,d,U^\lambda}) + q_1(|D_t^{m,d,U^\lambda}|)\Big)\Big).$$

As $t \mapsto |D_t^{m,d,U^\lambda}|$ is decreasing, it follows that

$$V(m,d) \leq V^{\tilde{U}^{\lambda,n}}(m,d) + C\Big(\frac{1}{1-\beta}\Big)\Big(\sup_m \lambda(m,d) + q_1(|d|)\Big).$$

By taking $n \to \infty$, the result follows. □

Appendix B

Interim results for Robust Gittins index theorem and a computation for a robust Bernoulli bandit

This appendix presents interim results for the proof of the robust Gittins' theorem in Chapter 8 and the algorithm to compute the robust index for the Bernoulli case considered in Chapter 7.

B.1 Interim results

Definition B.1 (Föllmer and Schied [49])**.** Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and let $\mathcal{Y}$ be a family of $\mathcal{G}$-measurable random variable. We say Z is a $\mathcal{G}$-*essential infimum* of $\mathcal{Y}$ denoted by $Z = \mathcal{G}\text{-}\operatorname{ess\,inf} Y$ if

(i) Z is $\mathcal{G}$-measurable.

(ii) $Z \leq Y \ \ \mathbb{P}-a.s.$ for all $Y \in \mathcal{Y}$.

(iii) For Z' such that $Z' \leq Y \ \ \mathbb{P}-a.s.$ for all $Y \in \mathcal{Y}$, we must have $Z' \leq Z \ \mathbb{P}-a.s.$.

We also define a similar notion for $\mathcal{G}$-*essential supremum*. We may omit $\mathcal{G}$ in front of ess inf if the measurability of the family is obvious.

Theorem B.1 (Existence of Essential infimum)**.** *The $\mathcal{G}$-essential infimum exists.*

Suppose in addition that $\mathcal{Y}$ is directed downwards, that is for $Y, Y' \in \mathcal{Y}$, there exists $\tilde{Y} \in \mathcal{Y}$ such that $\tilde{Y} \leq \min(Y, Y')$. Then there exists a decreasing sequence $(Y_n)_{n\in\mathbb{N}} \subseteq \mathcal{Y}$ such that $Y_n \searrow \operatorname{ess\,inf} Y \ \ \mathbb{P}$-a.s.

A similar result also holds for $\mathcal{G}$-essential supremum.

Lemma B.2. *Let $V_s : \mathcal{T}(s) \to L^\infty(\mathcal{F}_s)$ be a function such that for every $\tau, \sigma \in \mathcal{T}(s)$ and $A \in \mathcal{F}_s$, we have $V_s(\tau\, \mathbb{I}_A + \sigma\, \mathbb{I}_{A^c}) = V_s(\tau)\mathbb{I}_A + V_s(\sigma)\mathbb{I}_{A^c}$.*

Then there exists a sequence $\tau_n \in \mathcal{T}(s)$ such that $V_s(\tau_n) \searrow \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} V_s(\tau)$ $\mathbb{P}$-a.s..

Proof. For $\tau, \sigma \in \mathcal{T}(s)$, we define $A := \{V_s(\tau) > V_s(\sigma)\} \in \mathcal{F}_s$. Then

$$\tilde{\tau} := \tau\, \mathbb{I}_A + \sigma\, \mathbb{I}_{A^c} \in \mathcal{T}(s).$$

By assumption, $V_s(\tilde{\tau}) \geq V_s(\tau) \wedge V_s(\sigma)$; the result follows from Theorem B.1. □

Lemma B.3. *Let $f : \mathcal{T}(s) \times L^\infty(\mathcal{F}_s) \to L^\infty(\mathcal{F}_s)$ satisfy:*

(i) For all $\tau \in \mathcal{T}(s)$, $f(\tau, 0) \geq 0$ $\mathbb{P}$-a.s. and $f(\tau, X) < 0$ for some $X \in L^\infty(\mathcal{F}_s)$.

(ii) There exists $L \in [0, \infty)$ such that, for every $X, Y \in L^\infty(\mathcal{F}_s)$ with $X \geq Y$ $\mathbb{P}$-a.s and $\tau \in \mathcal{T}(s)$, we have $0 \leq f(\tau, Y) - f(\tau, X) \leq L(X - Y)$ $\mathbb{P} - a.s.$

(iii) For all $A \in \mathcal{F}_s$, all $\tau, \sigma \in \mathcal{T}(s)$ and all $X, Y \in L^\infty(\mathcal{F}_s)$, we have

$$f(\tau \mathbb{I}_A + \sigma \mathbb{I}_{A^c}, X\mathbb{I}_A + Y\mathbb{I}_{A^c}) = f(\tau, X)\mathbb{I}_A + f(\sigma, Y)\mathbb{I}_{A^c}.$$

Define $X^* := \operatorname{ess\,inf} \left\{ X \in L^\infty(\mathcal{F}_s) \;:\; \operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X) \leq 0 \;\; \mathbb{P}\text{-a.s.} \right\}$.
Then $X^* \in L^\infty(\mathcal{F}_s)$ and $\operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X^*) = 0$ $\mathbb{P}$-a.s.
Note: All essential infima in this lemma are taken among the $\mathcal{F}_s$-measurable functions.

Proof. Denote $\mathcal{X} := \left\{ X \in L^\infty(\mathcal{F}_s) \;:\; \operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X) \leq 0 \;\; \mathbb{P}\text{-a.s.} \right\}$.

By (i), $\mathcal{X} \neq \phi$. For a fixed $X \in \mathcal{X}$, by Lemma B.2 there exists a sequence $\tau_k \in \mathcal{T}(s)$, such that $f(\tau_k, X) \searrow \operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X)$ $\mathbb{P}$-a.s. Similarly, we can find a sequence (τ'_k) for $X' \in \mathcal{X}$.

Define a sequence $\sigma_k := \tau_k \mathbb{I}_A + \tau'_k \mathbb{I}_{A^c}$ where $A = \{X \leq X'\}$. Then

$$\begin{aligned}
&\operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f\left(\tau, \min(X, X')\right) \leq f(\sigma_n, \min(X, X')) \\
&\quad = f(\tau_k \mathbb{I}_A + \tau'_k \mathbb{I}_{A^c}, X\mathbb{I}_A + X'\mathbb{I}_{A^c}) = f(\tau_k, X)\mathbb{I}_A + f(\tau'_k, X')\mathbb{I}_{A^c} \\
&\quad \searrow \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X)\mathbb{I}_A + \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X')\mathbb{I}_{A^c} \;\; \leq \;\; 0.
\end{aligned}$$

Hence, $\mathcal{X}$ is downward directed. Therefore, by Theorem B.1, there exists a sequence $(X_n)_{n \geq 0} \subseteq \mathcal{X}$ such that $X_n \searrow X^*$ $\mathbb{P}$-a.s. This implies that X^* is almost surely bounded from above. By monotonicity, as $f(\tau, 0) \geq 0$, it follows from strict monotonicity that $f(\tau, -1) > 0$. Therefore, -1 is an essential lower bound of $\mathcal{X}$, so X^* is bounded below by -1 and $X^* \in L^\infty(\mathcal{F}_s)$.

For the final assertion, we will first show that $\operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X^*) \leq 0$ $\mathbb{P}$-a.s. For each $n \in \mathbb{N}$, we can again find a sequence τ_k^n such that

$$f(\tau_k^n, X_n) \searrow \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X_n) \leq 0 \quad \text{as} \quad k \to \infty \quad \mathbb{P}\text{-a.s..}$$

By condition (ii), it follows that

$$\begin{aligned}
L(X_n - X^*) &\geq f(\tau_k^n, X^*) - f(\tau_k^n, X_n) \geq \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X^*) - f(\tau_k^n, X_n) \\
&\nearrow \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X^*) - \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X_n) \qquad \text{as} \quad k \to \infty.
\end{aligned}$$

Hence, $L(X_n - X^*) \geq \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*)$.

By taking $n \to \infty$, it follows that $\operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*) \leq 0$.

To finish the proof, it suffices to show that for all $\sigma \in \mathcal{T}(s)$, we have $f(\sigma, X^*) \geq 0$. Fix $\sigma \in \mathcal{T}(s)$. Define $F := -f(\sigma, X^*)$ and the event $B := \{F > 0\}$. Let $\tau_k \in \mathcal{T}(s)$ be a sequence such that $f(\tau_k, X^*) \searrow \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*) \leq 0$. We define $Y := X^*\mathbb{I}_{B^c} + \left(X^* - \frac{F}{2L}\right)\mathbb{I}_B \leq X^*$ and a sequence $\sigma_k = \tau_k\mathbb{I}_{B^c} + \sigma\mathbb{I}_B$. Then

$$\begin{aligned}
\operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, Y) &\leq f(\sigma_k, Y) = f(\tau_k, X^*)\mathbb{I}_{B^c} + f\Big(\sigma, X^* - \frac{F}{2L}\Big)\mathbb{I}_B \\
&= f(\tau_k, X^*)\mathbb{I}_{B^c} + \left(\left(f\Big(\sigma, X^* - \frac{F}{2L}\Big) - f\left(\sigma, X^*\right)\right) - F\right)\mathbb{I}_B \\
&\leq f(\tau_k, X^*)\mathbb{I}_{B^c} + \left(L\left(X^* - \left(X^* - \frac{F}{2L}\right)\right) - F\right)\mathbb{I}_B \\
&= f(\tau_k, X^*)\mathbb{I}_{B^c} - \frac{F}{2}\mathbb{I}_B \leq f(\tau_k, X^*)\mathbb{I}_{B^c} \quad \searrow \quad \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*)\mathbb{I}_{B^c} \leq 0.
\end{aligned}$$

Hence, $Y \in \mathcal{X}$. By minimality of X^*, it must follow that B is a $\mathbb{P}$-null set. □

Theorem B.4. *There exists $\tau^* \in \mathcal{T}(s)$ such that $V_s(\tau^*, \lambda) = \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} V_s(\tau, \lambda)$.*

Proof. We write $\mathcal{F}_n^s := \mathcal{F}_{s+n}$ and define the processes

$$Y_0 := \frac{(C+1) - \lambda}{1-\beta}, \quad Y_n := \sum_{t=s+1}^{s+n} \beta^t \left(h(t) - \lambda\right) \quad \text{and} \quad Z_n := \operatorname*{ess\,inf}_{\tau\geq n} \mathcal{E}\left(Y_\tau \middle| \mathcal{F}_n^s\right)$$

where τ is considered over the space of all stopping times and C is an upper bound given in Assumption 6.2.

Define $\tau^* := \inf\{n \geq 0 : Y_n = Z_n\}$.

By robust representation theorem (Theorem 6.3), we can represent $\mathcal{E}$ as an essential supremum over a familiy of probability measures which satisfy the law of iteration ([90, Equation 4]). It then follows Riedel [90, Theorem 3] that if $\tau^* < \infty$ $\mathbb{P}$-a.s., then τ^* is an optimal solution.

Hence, to prove the required result, it suffices to prove that $\tau^* \in \mathcal{T}(s)$, $Z_0 = \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} V_s(\tau, \lambda)$ and $\tau^* < \infty$ $\mathbb{P}$-a.s.

It is clear that we never stop at time 0. Thus, $\tau^* \in \mathcal{T}(s)$ and $Z_0 = \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} V_s(\tau, \lambda)$.

On an event ω such that $\lambda(\omega) < C$ and $h(t)(\omega) \to C$, we can find $N(\omega)$ sufficiently large such that $\sum_{t=N+1}^{n} \beta^t \left(h(s+t) - \lambda\right)(\omega) > 0$ for all n. In particular, we have $\tau^*(\omega) \leq N(\omega) < \infty$. Therefore, we have $\tau^* < \infty$ $\mathbb{P}$-a.s. and thus it must be optimal. □

Lemma B.5. *Let $(\rho_n)_{0\leq n\leq L-1}$ be a simple form of (τ, p). Then the sequence $(\rho_n)_{0\leq n\leq L-1}$ can be expressed recursively by the following relation. Set $\rho_0 = p_0$ and define*

$$\rho_n = \begin{cases} \rho_{n-1} & \text{if } \sum_{m=1}^K \pi_n^{(m)} \geq n, \\ p_{\psi_n} & \text{if } \sum_{m=1}^K \pi_n^{(m)} = n-1, \end{cases} \quad \text{where} \quad \pi_n^{(m)} := \sum_{i=0}^{\hat{F}_n^{(m)}-1} \tau_i^{(m)}, \tag{B.1.1}$$

and where

$$\psi_n = \sum_{m=1}^K \hat{F}_n^{(m)} \quad \text{and} \quad \hat{F}_n^{(m)} := \min\left\{ f \geq 0 \ : \ \sum_{i=0}^{n-1} \mathbb{I}(\rho_i = m) \leq \sum_{i=0}^{f-1} \tau_i^{(m)} \right\}.$$

Proof. To see this, we view $\hat{F}_n^{(m)}$ as the number of runs (under (τ, p)) of the mth bandit before making the nth play. We then consider $\pi_n^{(m)}$ as the total number of trials in the mth bandit required to complete the $\hat{F}_n^{(m)}$th run.

If the run is not yet completed before the nth play, we continue to play the same machine. (i.e. we define $\rho_n = \rho_{n-1}$). If the run is completed in the $(n-1)$th play, we make a new decision in the nth play based on the choice sequence p. In that case, we have already made $\psi_n = \sum_{m=1}^{K} \hat{F}_n^{(m)}$ decisions before the nth play. By our convention to start at ρ_0, the decision of the nth play is given by p_{ψ_n}. □

Definition B.2. Given a simple form choice sequence ρ, we define the *decision filtration* induced by ρ by

$$\mathcal{G}_n^\rho := \{ A \in \mathcal{F}(T) \ : \ A \cap \{\tilde{\eta}_n = r\} \in \mathcal{F}(r) \ \} \tag{B.1.2}$$

where $\tilde{\eta}$ is the recording sequence corresponding to $(\mathbf{1}, \rho)$.

Lemma B.6. *The sequence of σ-algebras $(\mathcal{G}_n^\rho)$ given in Definition B.2 forms a filtration, i.e. $\mathcal{G}_n^\rho \subseteq \mathcal{G}_{n+1}^\rho$.*

Proof. Suppose that $A \in \mathcal{G}_n^\rho$, then for $r \in \mathcal{S}$ and $m \in \{0\} \cup [K]$,

$$A \cap \{\tilde{\eta}_n = r - e^{(m)}\} \in \mathcal{F}(r - e^{(m)}) \subseteq \mathcal{F}(r).$$

Moreover, by definition, $\{\rho_n = m\} \in \mathcal{G}_n^\rho$, it follows that

$$A \cap \{\tilde{\eta}_{n+1} = r\} = \bigcup_{m=0}^{K} \left(A \cap \{\tilde{\eta}_n = r - e^{(m)}\} \right) \cap \left(\{\rho_n = m\} \cap \{\tilde{\eta}_n = r - e^{(m)}\} \right) \in \mathcal{F}(r).$$

Hence, $A \in \mathcal{G}_{n+1}^\rho$. □

Lemma B.7. *With (ξ_n^ρ) as in Remark 7.6 and $(\mathcal{G}_n^\rho)$ in Definition B.2, we have ξ_n^ρ is $\mathcal{G}_n^\rho$-measurable. In particular, $\mathcal{H}_n^\rho := \sigma(\xi_1^\rho, ..., \xi_n^\rho) \subseteq \mathcal{G}_n^\rho$.*

Proof. By Lemma B.6, ρ_{n-1} is $\mathcal{G}_n^\rho$-measurable. Hence, for a fixed $B \in \mathcal{B}$, we have

$$\{\xi_n^\rho \in B\} \cap \{\tilde{\eta}_n = r\} = \left\{ \xi_{r^{(m)}}^{(m)} \in B \right\} \cap \underbrace{\{\tilde{\eta}_n = r\} \cap \{\rho_{n-1} = m\}}_{\in \mathcal{F}(r) \text{ as } \{\rho_{n-1}=m\} \in \mathcal{G}_n^\rho} \in \mathcal{F}(r).$$

Therefore, ξ_n^ρ is $\mathcal{G}_n^\rho$-measurable and the last assertion follows from Lemma B.6. □

Theorem B.8. *With $(\mathcal{H}_n^\rho)_{n\geq 0}$ as in Remark 7.6 and $(\mathcal{G}_n^\rho)_{n\geq 0}$ in Definition B.2, we have $\mathcal{G}_n^\rho = \mathcal{H}_n^\rho$.*

Proof. We will prove this result by induction. It is clear that $\mathcal{G}_0^\rho = \mathcal{H}_0^\rho$. We now assume that $\mathcal{G}_{n-1}^\rho = \mathcal{H}_{n-1}^\rho$.

By Lemma B.7, it suffices to show that $\mathcal{G}_n^\rho \subseteq \mathcal{H}_n^\rho$. Recall from Definition 7.3 that the recording sequence $\tilde{\eta}_n$, corresponding to a simple form, takes values in

$$\mathcal{S}_n := \left\{ r \in \mathcal{S} \ : \ \sum_{m=1}^{K} r^{(m)} = n \right\}.$$

For a fixed $r \in \mathcal{S}_n$, we define $\mathcal{P}_r$ to be the space of sequences of length $\sum_{m=1}^{K} r^{(m)}$ with values in $[K]$ with exactly $r^{(m)}$ replications of $m \in [K]$. (e.g. for $[K] = \{1,2\}$, $\mathcal{P}_{(3,1)} = \{(1,1,1,2),(1,1,2,1),(1,2,1,1),(2,1,1,1)\}$).

For $A \in \mathcal{G}_n^\rho$ and $\pi \in \mathcal{P}_r$,

$$A \cap \left\{(\rho_i)_{0\le i\le n-1} = \pi\right\} = A \cap \left\{(\rho_i)_{0\le i\le n-1} = \pi\right\} \cap \{\tilde{\eta}_n = r\} \in \mathcal{F}(r).$$

By the Doob–Dynkin lemma, there exists a measurable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\begin{aligned}\mathbb{I}_{A\cap\{(\rho_i)_{0\le i\le n-1}=\pi\}} &= f\Big(\big(\xi_t^{(m)}\big)_{1\le t\le r^{(m)},m\in[K]}\Big) f\Big(\big(\xi_t^{(m)}\big)_{1\le t\le r^{(m)},m\in[K]}\Big)\mathbb{I}_{\{(\rho_i)_{0\le i\le n-1}=\pi\}} \\ &= f_\pi\Big(\,(\xi_i^\rho)_{1\le i\le n}\,\Big)\mathbb{I}_{\{(\rho_i)_{0\le i\le n-1}=\pi\}} \qquad \text{(B.1.3)}\end{aligned}$$

for some measurable function f_π defined by reordering the input of f by π.

As $\mathbb{I}_{\{(\rho_i)_{0\le i\le n-1}=\pi\}}$ is $\mathcal{G}_{n-1}^\rho$-measurable, it must be $\mathcal{H}_{n-1}^\rho$-measurable by the induction hypothesis. Therefore, $\mathbb{I}_A = \sum_{r\in\mathcal{S}_N}\sum_{\pi\in\mathcal{P}_r}\mathbb{I}_{A\cap\{(\rho_i)_{0\le i\le n-1}=\pi\}}$ is $\mathcal{H}_n^\rho$-measurable by (B.1.3). This means that $A \in \mathcal{H}_n^\rho$, which completes the proof. □

B.2 Proofs of stated results

Proof of Theorem 7.1. We will omit superscript (k) for notational simplicity.

By Corollary 8.2 and Theorem 6.3 (together with positive homogeneity), we have that, for any $\tau \in \mathcal{T}(s)$,

$$\operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right) \ge 0. \qquad \text{(B.2.1)}$$

By Lemma 11.19 (together with the construction of Theorem 11.22) in Föllmer and Schied [49], the family $\left\{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\big|\mathcal{F}_s\big) : \mathbb{Q}\in\mathcal{Q}\right\}$ must be directed upwards. Hence, by Theorem B.1, we can find a family $(\mathbb{Q}_n) \subseteq \mathcal{Q}$ such that

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\right) \xrightarrow[n\to\infty]{} \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right). \qquad \text{(B.2.2)}$$

Let $\tilde{\Omega}$ be an event with probability one such that the followings hold.

1. (B.2.1) and (B.2.2) holds.
2. For all $n \in \mathbb{N}$,

$$\frac{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\big)} \le \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\big)},$$

$\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\big) \ge \beta$, and

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right) = \mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\middle|\mathcal{F}_s\right) - \gamma(s)\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\middle|\mathcal{F}_s\right). \qquad \text{(B.2.3)}$$

Fix $\omega \in \tilde{\Omega}$ and $\epsilon > 0$. By (B.2.1) and (B.2.2), there exists $n \in \mathbb{N}$ such that

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau} \beta^t \big(h(s+t) - \gamma(s)\big) \middle| \mathcal{F}_s \right)(\omega) \geq -\epsilon.$$

By (B.2.3), we can rearrange the inequality above and obtain

$$\begin{aligned}\gamma(s)(\omega) &\leq \frac{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)}(\omega) + \frac{\epsilon}{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)(\omega)} \\ &\leq \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)}(\omega) + \frac{\epsilon}{\beta}.\end{aligned}$$

As ϵ is arbitrary, it follows that on $\tilde{\Omega}$,

$$\gamma(s) \leq \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)}.$$

Hence,

$$\gamma(s) \leq \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)}.$$

By Theorem 8.6 and Theorem 6.3, we can find $\sigma := \sigma(s, \gamma(s)) \in \mathcal{T}(s)$ such that,

$$\operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma} \beta^t \big(h(s+t) - \gamma(s)\big) \middle| \mathcal{F}_s \right) = 0.$$

Hence, for all $\mathbb{Q} \in \mathcal{Q}$, $\mathbb{E}^{\mathbb{Q}}\Big(\sum_{t=1}^{\sigma} \beta^t \big(h(s+t) - \gamma(s)\big) \Big| \mathcal{F}_s \Big) \leq 0$. Therefore,

$$\gamma(s) \geq \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\sigma} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\sigma} \beta^t \big| \mathcal{F}_s \big)}.$$

In particular,

$$\gamma(s) \geq \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\sigma} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\sigma} \beta^t \big| \mathcal{F}_s \big)} \geq \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t) \big| \mathcal{F}_s \big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s \big)}.$$

This completes the proof. □

Proof of Proposition 6.4. (i) is straightforward.

By Theorem 6.3 and the tower property,

$$\begin{aligned}\operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\big(X \big| \mathcal{F}(S)\big) &= \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\mathbb{E}^{\mathbb{Q}}\big(X \big| \mathcal{F}(S')\big) \Big| \mathcal{F}(S)\Big) \\ &\leq \operatorname*{ess\,sup}_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\operatorname*{ess\,sup}_{\mathbb{Q}' \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}'}\big(X \big| \mathcal{F}(S')\big) \Big| \mathcal{F}(S)\Big).\end{aligned}$$

Hence, (ii) follows.

For $\hat{Y}(\omega^{(1)},...,\omega^{(K)}) = X^{(1)}(\omega_1)\times\cdots\times X^{(K)}(\omega_K)$, by Fubini's theorem,

$$\operatorname*{ess\,sup}_{\otimes_{m=1}^K \mathbb{Q}^{(m)}\in\mathcal{Q}} \mathbb{E}^{\otimes_{m=1}^K \mathbb{Q}^{(m)}}\left(\prod_{m=1}^K X^{(m)}\middle|\mathcal{F}(S)\right) = \operatorname*{ess\,sup}_{\otimes_{m=1}^K \mathbb{Q}^{(m)}\in\mathcal{Q}} \prod_{m=1}^K \mathbb{E}^{\mathbb{Q}^{(m)}}\left(X^{(m)}\middle|\mathcal{F}^{(m)}_{S^{(m)}}\right).$$

Then (iii) and (iv) follow by considering different choices of $X^{(m)}$. □

Proof of Theorem 8.1. This can be done by showing that V_s satisfies the regularity assumptions of Lemma B.3.

By considering $\lambda = C+1$, where C is an upper bound on h, we see that $V_s(\tau, C+1) < 0$. As $h(t)\geq 0$, it also follows that $V_s(\tau,0)\geq 0$. Hence, condition (i) is satisfied.

For condition (ii), suppose that $\lambda' > \lambda$. Then

$$\begin{aligned}\left(\sum_{t=s+1}^{s+\tau}\beta^t\big(h(t)-\lambda\big)\right) &= \left(\sum_{t=s+1}^{s+\tau}\beta^t\big(h(t)-\lambda'\big)\right) + \sum_{t=s+1}^{s+\tau}\beta^t(\lambda'-\lambda)\\ &\leq \left(\sum_{t=s+1}^{s+\tau}\beta^t\big(h(t)-\lambda'\big)\right) + \left(\frac{\beta^{s+1}}{1-\beta}\right)(\lambda'-\lambda).\end{aligned}$$

By monotonicity and translation equivariance, we have

$$\begin{aligned}0 \leq V_s(\tau,\lambda) - V_s(\tau,\lambda') &= \mathcal{E}\left(\sum_{t=s+1}^{s+\tau}\beta^t\big(h(t)-\lambda\big)\,\middle|\,\mathcal{F}_s\right) - \mathcal{E}\left(\sum_{t=s+1}^{s+\tau}\beta^t\big(h(t)-\lambda'\big)\,\middle|\,\mathcal{F}_s\right)\\ &\leq \left(\frac{\beta^{s+1}}{1-\beta}\right)(\lambda'-\lambda).\end{aligned}$$

So, V_s is Lipschitz in λ.

Condition (iii) follows from $(\mathcal{F}_t)_{t\geq 0}$-regularity of $\mathcal{E}$ (Remark 6.3). □

B.3 Numerical algorithm to estimate Robust Gittins indices

To approximate the value of $\gamma_{i,\beta,T}\left(p, 1/\sqrt{n}\right)$ in the setting of Section 7.3, we proceed following the rough recipe below.

Fix a grid of values for $\gamma\in G\subseteq[0,1]$.

1. Set $V_T^\gamma\left(p,\frac{1}{\sqrt{n+T}}\right) = 0$.

2. Assume that we know V_{t+1}^γ. Evaluate the backward recursion

$$V_t^\gamma\left(p,\frac{1}{\sqrt{n+t}}\right) = \min\left\{0, \mathcal{E}_{(t)}\left((h(t)-\gamma)+\beta V_{t+1}^\gamma\left(p_{t+1},\frac{1}{\sqrt{n+(t+1)}}\right)\right)\Big|_{(p_t,n_t)=(p,n+t)}\right\}$$

   This is done by considering the discrete values of p and using linear interpolation over $[0,1]$.

3. Using these iterates, determine the initial value function

$$U\Big(\gamma, p, \frac{1}{\sqrt{n+t}}, T-t\Big) = \mathcal{E}_{(t)}\Big((h(t+1)-\gamma) + \beta V_{t+1}^{\gamma}\Big(p_{t+1}, \frac{1}{\sqrt{n+t+1}}\Big)\Big)\Big|_{(p_t,n_t)=(p,n+t)}.$$

By the Snell's envelope argument, we obtain

$$U\Big(\gamma, p, \frac{1}{\sqrt{n+t}}, T-t\Big) = \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} \mathcal{E}\Big(\sum_{s=t+1}^{\tau} \beta^{s-t}\,(h(s)-\gamma)\,\Big|\,\mathcal{F}_t\Big)\Big|_{(p_t,n_t)=(p,n+t)}.$$

4. Repeat step 1-3 to compute $U\big(\gamma, p, \frac{1}{\sqrt{n+t}}, T-t\big)$ for all $\gamma \in G$.

5. Calculate $\gamma_{\alpha,\beta,T-t}\Big(p, \frac{1}{\sqrt{n+t}}\Big)$ for a fixed $(p, 1/\sqrt{n+t})$ by

$$\gamma_{\alpha,\beta,T-t}\Big(p, \frac{1}{\sqrt{n+t}}\Big) = \min\Big\{\gamma \in G : U\big(\gamma, p, \frac{1}{\sqrt{n+t}}, T-t\big) \le 0\Big\}.$$

Bibliography

[1] R. Agrawal. Sample mean based index policies by $O(\log N)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

[2] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *JMLR: Workshop and Conference Proceedings*, pages 1–26, 2012.

[3] L. Alili, P. Patie, and J. L. Pedersen. Representations of the First Hitting Time Density of an Ornstein-Uhlenbeck Process. *Stochastic Models*, pages 967–980, 2007.

[4] A. L. Allan and S. N. Cohen. Parameter Uncertainty in the Kalman–Bucy Filter. *SIAM Journal on Control and Optimization*, page 1646–1671, 2019.

[5] F. Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, pages 365–383, 1963.

[6] P. Armitage. *Sequential medical trials.* Blackwell Scientific, 1960.

[7] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, pages 203–228, 1999.

[8] P. Artzner, F. Delbaen, J. M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, pages 5–22, 2007.

[9] J. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 217–226, 2009.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, page 235–256, 2002.

[11] P. Bank and N. El Karoui. A stochastic representation theorem with applications to optimization and obstacle problems. *The Annals of Probability*, pages 1030–1067, 2004.

[12] P. Bank and H. Föllmer. American options, multi–armed bandits, and optimal consumption plans: A unifying view. *Paris-Princeton Lectures on Mathematical Finance 2002*, pages 1–42, 2002.

[13] P. Bank and C. Küchler. On Gittins' index theorem in continuous time. *Stochastic Processes and their Applications*, pages 1357–1371, 2007.

[14] J. A. Bather. *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, chapter Optimal stopping of Brownian motion: a comparison technique, pages 19–49. Academic Press, 1983.

[15] D. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Chapman & Hall, 1985.

[16] T. Bielecki, T. Chen, and I. Cialenco. Recursive construction of confidence regions. *Electronic Journal of Statistics*, pages 4674–4700, 2017.

[17] J. Bion-Nadal. Dynamic Risk Measures: Time Consistency and Risk Measures from BMO Martingales. *Finance and Stochastics*, pages 219–244, 2008.

[18] T. Björk, M. Khapko, and A. Murgoci. On time-inconsistent stochastic control in continuous time. *Finance Stoch.*, pages 331–360, 2017.

[19] T. Björk and A. Murgoci. A theory of Markovian time-inconsistent stochastic control in discrete time. *Finance Stoch.*, page 545–592, 2014.

[20] L. Breiman. First exit times from a square root boundary. In *Fifth Berkeley Symposium*, pages 9–16, 1967.

[21] M. Brezzi and T.L. Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, pages 87–108, 2002.

[22] G. Burtini, J. Loeppky, and R. Lawrence. A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit. *arXiv:1510.00757v4*, 2015.

[23] F. Caro and A. D. Gupta. Robust control of the multi-armed bandit problem. *Annals of Operations Research*, pages 1–20, 2015.

[24] J. Chakravirty and A. Mahajan. Multi-armed bandits, gittins index and its calculation. *Methods and Applications of Statistics in Clinical Trials: Planning Analysis and Inferential Methods*, pages 416–435, 2014.

[25] F. Chang and T. L. Lai. Optimal stopping and dynamic allocation. *Advances in Applied Probability*, pages 829–853, 1987.

[26] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems 24*, page 2249–2257, 2011.

[27] H. Chernoff and A.J. Petkau. Numerical solutions for bayes sequential decision problems. *SIAM Journal of Scientic and Statistical Computing*, pages 46–59, 1986.

[28] S. N. Cohen. What risk measures are time consistent for all filtrations? *arXiv:1007.0610*, 2010.

[29] S. N. Cohen. Data-driven nonlinear expectations for statistical uncertainty in decisions. *Electronic Journal of Statistics*, pages 1858–1889, 2016.

[30] S. N. Cohen. Representing filtration consistent nonlinear expectations as g-expectations in general probability spaces. *Stochastic Processes and their Applications*, pages 1601–1626, 2018.

[31] S. N. Cohen and R. J. Elliott. A general theory of finite state Backward Stochastic Difference Equations. *Stochastic Processes and their Applications*, pages 442–466, 2010.

[32] S. N. Cohen and R. J. Elliott. Backward Stochastic Difference Equations and nearly-time-consistent nonlinear expectations. *SIAM Journal on Control and Optimization*, pages 125–139, 2011.

[33] S. N. Cohen and R. J. Elliott. *Stochastic Calculus and Applications*. Birkhäuser, 2015.

[34] S. N. Cohen and T. Treetanthiploet. Asymptotic Randomised Control with applications to bandits. arXiv:2010.07252, 2020.

[35] S. N. Cohen and T. Treetanthiploet. Gittins' theorem under uncertainty. arXiv:1907.05689, 2020.

[36] S. N. Cohen and T. Treetanthiploet. Correlated bandits for dynamic pricing via the arc algorithm. arXiv:2102.04263, 2021.

[37] F. Coquet, Y. Hu, J. Mémin, and S. Peng. Filtration consistent nonlinear expectations and related g-expectations. *Probability Theory and Related Fields*, pages 1–27, 2002.

[38] W. Cowan, J. Honda, and M. N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, pages 1–28, 2018.

[39] K. Detlefsen and G. Scandolo. Conditional and dynamic convex risk measures. *Finance Stochastics*, pages 539–561, 2005.

[40] J-P. Dubé and S. Misra. Scalable price targeting. NBER working paper No. w23775, National Bureau of Economic' Research, 2017.

[41] I. Ekren, N. Touzi, and J. Zhang. Optimal stopping under nonlinear expectation. *Stochastic Processes and their Applications*, pages 3277–3311, 2014.

[42] N. El Karoui and I. Karatzas. Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, pages 255–286, 1994.

[43] N. El Karoui, S. Peng, and M. C. Quenez. Backward Stochastic Differential Equations in finance. *Mathematical Finance*, pages 1–71, 1997.

[44] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, page 643–669, 1961.

[45] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, page 1079–1105, 2006.

[46] L. Fahrmeir. Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, pages 501–509, 1992.

[47] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: the Generalized Linear case. In *NIPS'10: Proceedings of the 23rd International Conference on Neural Information Processing Systems*, page 586–594, 2010.

[48] H. Föllmer and I. Penner. Convex risk measures and the dynamics of their penalty functions. *Statistics & Decisions*, pages 61–96, 2006.

[49] H. Föllmer and A. Schied. *Stochastic Finance: an introduction in discrete time*. De Gruyler, 2016.

[50] M. Frittelli and E. R. Gianin. Putting order in risk measures. *Journal of Banking & Finance*, pages 1473–1486, 2002.

[51] M. Frittelli and G. Scandolo. Risk measures and capital requirements for processes. *Mathematical Finance*, pages 589–612, 2006.

[52] E. Frostig and G. Weiss. Four proofs of Gittins' multiarmed bandit theorem. *Annals of Operations Research*, pages 127–165, 2016.

[53] R. Fryer and P. Harms. Two-armed restless bandits with imperfect information: Stochastic control and indexability. *Mathematics of Operations Research*, pages 399–427, 2018.

[54] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266, Amsterdam: North Holland, 1974.

[55] J. C. Gittins and D. M. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, pages 561–565, 1979.

[56] J.C. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley and Sons, 1989.

[57] S. Graf. A Radon–Nikodym theorem for capacities. *für die reine und angewandte Mathematik*, pages 192–214, 2009.

[58] Y. Hu, H. Jin, and X. Y. Zhou. Time-inconsistent stochastic linear-quadratic control. *SIAM J. Control and Optimization*, pages 1548–1572, 2012.

[59] X. Huo and F. Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. 2017.

[60] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, pages 257–280, 2005.

[61] D. Kahneman and A. Tversky. Prospect theory: Analysis of decision under risk. *Econometrica*, pages 263–292, 1979.

[62] I. Karatzas. Gittins indices in the dynamic allocation problem for diffusion processes. *Annals of Probability*, pages 173–192, 1984.

[63] N. El Karoui and I. Karatzas. General Gittins index processes in discrete time. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1232–1236, 1993.

[64] E. Kaufmann, O. Cappé, and Aur´elien Garivier. On Bayesian Upper Confidence Bounds for Bandit problems. In *Artificial intelligence and statistics*, pages 592–600, 2012.

[65] J. M. Keynes. *A Treatise on Probability*. Macmillan and Co., 1921. Reprint BN Publishing, 2008.

[66] M. J. Kim and A. E.B. Lim. Robust multiarmed bandit problems. *Management Science*, pages 264–285, 2015.

[67] J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. *Proceedings of Machine Learning Research*, pages 1–28, 2018.

[68] F. H. Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin, 1921. reprint Dover 2006.

[69] T. C. Koopmans. Stationary ordinal utility and impatience. *Econometrica*, pages 287–309, 1960.

[70] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, pages 4–22, 1985.

[71] T. Lattimore. Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. *arXiv:1511.06014*, 2015.

[72] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

[73] N. N. Lebedev. *Special functions and their applications*. Dover Publications, 1972.

[74] J. Li. The k-armed bandit problem with multiple priors. *Journal of Mathematical Economics*, pages 22–38, 2019.

[75] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, pages 661–670, 2010.

[76] O. A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *In Conference on Learning Theory*, pages 497–514, 2011.

[77] A. Mandelbaum. Discrete multi-armed bandits and multi-parameter processes. *Probabability Theory and Related Fields*, pages 129–147, 1986.

[78] A. Mandelbaum. Continuous multi-armed bandits and multi-parameter processes. *The Annals of Applied Probability*, pages 1527–1556, 1987.

[79] K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, pages 226–252, 2019.

[80] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, pages 780–798, 2005.

[81] B. Peleg and M. E. Yaari. On the Existence of a Consistent Course of Action when Tastes are Changing. *The Review of Economic Studies*, pages 391–401, 1973.

[82] S. Peng. *Backward Stochastic Differential Equations, chapter 9: Backward SDE and related g-expectation.* Pitman Research Notes in Mathematics, Longman, 1997. 141-159.

[83] S. Peng. Nonlinear Expectations and Stochastic Calculus under uncertainty, 2010.

[84] H. Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications.* Springer, 2009.

[85] R. A. Pollak. Consistent planning. *The Review of Economic Studies*, pages 201–208, 1968.

[86] J. Quiggin. A theory of anticipated utility. *Journal of Economic Behavior and Organization*, pages 323–343, 1982.

[87] J. Quiggin. *Generalized Expected Utility Theory. The Rank-Dependent Model.* Kluwer Academic, Boston, 1993.

[88] C. Reisinger and Y. Zhang. Regularity and stability of feedback relaxed control. *arXiv:2001.03148*, 2020.

[89] F. Riedel. Dynamic coherent risk measures. *Stochastic Processes and their Applications*, pages 185–200, 2004.

[90] F. Riedel. Optimal Stopping With Multiple Priors. *Econometrica*, pages 857–908, 2009.

[91] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, pages 527–535, 1952.

[92] R. Rockafellar. *Convex Analysis*. Princeton university press, 1972.

[93] P. Rusmevichientong, A.J. Mersereau, and J. N. Tsitsiklis. A Structured Multiarmed Bandit Problem and the Greedy Policy. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2787 – 2802, 2009.

[94] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, pages 395–411, 2009.

[95] D. Russo. A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *arxiv.1904.04732*, 2019.

[96] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, pages 1–30, 2016.

[97] D. Russo and B. Van Roy. Learning to Optimize via Information-Directed Sampling. *Operational Research*, pages 1–23, 2017.

[98] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, pages 1–96, 2018.

[99] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, pages 235–261, 2010.

[100] I. O. Ryzhov, W. B. Powell, and P. I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 2012.

[101] W. Shen, J. Wang, Y. G. Jiang, and H. Zha. Portfolio choices with orthogonal bandit learning. *Proceeding IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence*, pages 974–980, 2015.

[102] R. H. Strotz. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, pages 165–180, 1955 - 1956.

[103] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. The MIT press, 2 edition, 2015.

[104] W. R. Thompson. on the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[105] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European Conference on Machine Learning.*, pages 437–448, 2005.

[106] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

[107] H. Wang, T. Zariphopoulou, and X. Zhou. Exploration versus exploitation in reinforcement learning: a stochastic control approach. *arXiv:1812.01552*, 2020.

[108] R. Weber. On the Gittins index for multi-armed bandits. *The Annals of Applied Probability*, pages 1024–1033, 1980.

[109] R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, pages 637–648, 1990.

[110] P. Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B*, pages 143–149, 1980.

[111] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, pages 287–298, 1988.

[112] Y. Yao. Some results on the Gittins index for a normal reward process. *Lecture Notes–Monograph Series*, pages 284–294, 2006.

[113] J. Yong. Time-inconsistent optimal control problems and the equilibrium HJB equation. *Mathematical Control & Related Fields*, pages 271–329, 2012.