Electronic Journal of Probability

Electron. J. Probab. **27** (2022), article no. 17, 1–48.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP742

Gittins' theorem under uncertainty

Samuel N. Cohen[*] [†] Tanut Treetanthiploet[‡] [§]

Abstract

We study dynamic allocation problems for discrete time multi-armed bandits under uncertainty, based on the the theory of nonlinear expectations. We show that, under independence assumption on the bandits and with some relaxation in the definition of optimality, a Gittins allocation index gives optimal choices. This involves studying the interaction of our uncertainty with controls which determine the filtration. We also run a simple numerical example which illustrates the interaction between the willingness to explore and uncertainty aversion of the agent when making decisions.

Keywords: Gittins index; uncertainty; nonlinear expectation; multi-armed bandits; time-consistency; robustness.
MSC2020 subject classifications: 93E35; 60G40; 91B32; 91B70.
Submitted to EJP on August 21, 2020, final version accepted on January 11, 2022.
Supersedes arXiv:1907.05689v3.

1 Introduction

When making decisions, people generally have a strict preference for options which they understand well. Since the classical work of Knight [52] and Keynes [50], there has been a stream of thinking 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.

This raises a natural question: how should we make decisions, given they will affect both our short-term outcomes, and the information available in the future? Shall we make a decision to explore and obtain new information, or shall we exploit the information available to optimize our profit? A simple setting in which this arises is a multi-armed bandit problem.

[*]Mathematical Institute, University of Oxford.
[†]The Alan Turing Institute, London.
E-mail: samuel.cohen@maths.ox.ac.uk
[‡]Mathematical Institute, University of Oxford.
[§]The Alan Turing Institute, London.
E-mail: ttreetanthiploet@turing.ac.uk

Modeling learning of the distribution of outcomes leads us to a paradox due to inconsistency in our decisions. As Keynes is said to have remarked[1], "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?"

In this paper, we use the theory of nonlinear expectations (or equivalently risk measures) which are known to model Knightian uncertainty (see Föllmer and Schied [36]). This has been used to address statistical uncertainty, for example in [23, 25]. To achieve consistency in decision making, we usually have to consider a time-consistent nonlinear expectation. This is widely studied through backward stochastic differential equations (BSDEs) (see, for example, the work of Peng and others [62, 61, 63, 59]).

However, these approaches presume that the flow of information is not controlled. (Formally, the filtration of our agent is fixed and independent of their controls.) When we can control the observations which we will receive, this is not the case. In order to address this issue, while accounting for uncertainty, we discuss an alternative approach to deriving a time-consistent control problem, based on ideas from indifference pricing and the martingale optimality principle. Using this approach, we show that when comparing different independent options, we can calculate an index separately for each alternative such that the 'optimal' strategy is always to choose the option with the smallest index. This idea was initially proposed by Gittins and Jones [41] (see also [42, 40]) in a context where the probability measure is fixed but estimation (in a Bayesian perspective) is modeled by the evolution of a Markov process.

Given this result, we demonstrate a numerical solution in a simple setting. We shall see that our algorithm gives behaviour which is both optimistic and pessimistic in different regimes, and compares well with existing methods for multi-armed bandits.

1.1 Multi-armed bandits

Multi-armed bandits are a classical toy example with which to study decision making with randomness. They are commonly known to have applications in medical trials (Armitage [4] or Anscombe [3]) and experimental design (Berry and Fristedt [13] or the classic paper of Robbins [69]), along with other areas. A few recent works in finance for portfolio selection can also be found in Huo and Fu [45] or Shen et al. [72]. The basic idea is that one has M 'bandits[2]', or equivalently, a bandit with M arms, and one must choose which bandit should be played at each time. A key paper studying these systems, Gittins and Jones [41], argued that for a collection of independent bandits, each governed by a countable state Markov process, one could compute the "Gittins index" for each bandit separately, and the optimal strategy is to play the machine with the lowest index (or the highest, depending on the sign of gains/losses). The proof of this result has been obtained using a number of different perspectives, for example Weber's prevailing charge formulation [75] (which we consider in more detail below), Whittle's retirement option formulation [76] and its extension without a Markov assumption by El Karoui and Karatzas [32] (and [33] in continuous time). A review of the proofs in discrete time is given by Frostig and Weiss [39]. However, in all these cases, the objective to be optimized is the discounted expected gain/loss – in particular, we are assumed to have no risk-aversion or uncertainty-aversion.

Gittins' index theory is commonly known as the first solution to an adaptive and

[1] It appears this quote may be misattributed. One suggestion (discussed by John Kay [48]) is that the correct attribution is to Paul Samuelson, and should read "When my information changes, I alter my conclusions", which fits even more easily with the thrust of this paper.

[2] 'One-armed bandits' are an early variety of automated gambling system, the descendants of which are also known as slot machines, fruit machines, or poker machines, depending on nationality.

sequential experimental design problem (from a Bayesian perspective) where the payoff of each bandit is assumed to be generated from a fixed unknown distribution[3] which must be inferred 'on-the-fly', but where experimentation may be costly. As an alternative to Gittins' index, Agrawal [1] proposed the 'Upper Confidence Bound (UCB) algorithm' which achieves a regret (deviation of average reward from the optimal reward) with the minimal asymptotic order of $\log(N)$, as proved by Lai and Robbins [54]. In the UCB algorithm, we compute a confidence interval for the expected reward at each step, and then play the bandit with the largest upper bound (where positive outcomes are preferred). Intuitively, using an upper bound encourages us to try bandits where we are less certain of the average reward, which encourages exploration. This is a form of 'optimism' in decisions, which is counter-intuitive from the classical 'pessimistic' utility theory (*à la* von Neumann and Morgenstern [74]), where our preferences are for more certain outcomes.

Typically, under appropriate assumptions, it is also the case that Gittins' index is a form of upper confidence bound for the estimated reward, an idea originally based on observations in Bather [12] and Kelly [49] and explored in more detail by Chang and Lai [21], followed up by Brezzi and Lai [18] (see Yao [78] for an error correction). Lattimore [55] also proves that Gittins' index achieves a minimal order bound on regret.

The apparent contradiction between the optimism of the UCB algorithm and Gittins' index and the pessimism of classical utility theory is what led to this paper. We extend the notion of Gittins' index to a robust (nonlinear) operator, allowing for uncertainty aversion. We work in a generic discrete-time setting, allowing for the possibility of online learning, non-stationary and continuous outcomes, embedding all these effects in an abstract 'nonlinear expectation'. (A concrete application to a simple setting with learning and uncertainty is given in Section 6.) In particular, we reformulate the proof of Gittins index theorem proposed by Weber [75], as this proof relies the least on the linearity of the expectation, and gives a natural form of time-consistency. We also remove a Markov assumption in Weber's proof by adapting El Karoui and Karatzas' formulation [32]. Our solution involves an optimal stopping problem under a nonlinear expectation, which can be converted to a low dimensional reflected BSDE (see for example El Karoui et al. [31] and Cheng and Riedel [22] in continuous time or An, Cohen and Ji [2] in discrete time). This allows us to see a balance between the desire to explore and to exploit in our decision making.

The robust version of Gittins index has some correlation to the adversarial bandit problem (see, for example, Auer et al. [7, 8]) where we are playing the bandit against an adversary. Our theory proposes an 'optimal' deterministic strategy (no additional randomness is introduced at the decision time) against an adversary who tries to maximize our cost, which is slightly different from the known random algorithms for the adversarial bandit problem. The key difference is that, in the classical adversarial problem, an adversary is trying to maximize our 'regret' whereas in our setting, we view the adversary as trying to maximize our cost. In our setting, the adversary is also permitted to respond to our current controls at every time, and we do not assume a minimax theorem holds.

The study of an adversary for the payoff in the bandit problem (via Gittins index) has been considered by Caro and Gupta [20] and Kim and Lim [51] (with additional penalty in the reward) using Whittle's retirement option argument [76]. In their works, they rely heavily on a Markov assumption, which allows them to postulate a robust dynamic programming principle (see also Iyengar [46], Nilim and El Ghaoui [58]). Their formulation considers the robust Gittins' strategy as a promising solution due to its

[3]There are a few variations on these assumption e.g. adversarial bandits, contextual bandits or non-stationary bandits. Reviews of these can be found in Burtini et al. [19] and Zhou [80].

optimality for a single bandit, but they do not show optimality for multiple bandits. Furthermore, their Markov assumption restricts them to have a fixed uncertainty at all times and, therefore, it is not clear how to incorporate learning in their model. In contrast, our framework pays more attention to defining a good notion of dynamic optimality for our nonlinear expectation without any Markov assumption.

By encoding learning through nonlinear expectations, a wide range of modeling options are included in our approach. For example, statistical concerns are treated in this framework in [23, 24] or Bielecki, Cialenco and Chen [14]. We could also allow adversarial choices with a range of a fixed set (as in the classical adversarial bandit problem [7] or as in [20, 51]) or a random set which can be used to model learning as in the classical Gittins' theory. We also allow dynamic adversaries, which are not considered in the usual adversarial setting.

The paper proceeds as follows: In Section 2, we present some relevant existing approaches to multi-armed bandits, which we will adapt and combine to obtain our result. In Section 3, we give the required definitions for the nonlinear expectations that we use to evaluate our decisions. We also discuss the different notions of optimality which are available, and how they interact with the dynamic programming principle.

In Section 4, we give a summary of how we apply these expectations to a multi-armed bandit problem, state the key result, and give a sketch outline of the proof. The full details of this (rather technical) proof are given in two appendices: Appendix A works through the first half of the proof, giving careful analysis of an optimal stopping problem under nonlinear expectation, and the corresponding 'fair value' process, for a single bandit; Appendix B gives the second half of the proof, and demonstrates that the single bandit analysis yields an optimal strategy when deciding between multiple bandits. Further technical lemmas, which are used but do not contributed significantly to the main proof, are given in Appendix C.

Section 5 considers a simple example of a multiple bandit problem numerically, suggesting some connections with behavioural finance; the algorithm used to compute this example is given in Appendix D.

2 Problem formulation and related approaches

2.1 General problem formulation

Broadly speaking, Gittins and Jones [41] argue that, in order to dynamically allocate a single resource amongst several alternative projects, the optimal policy is to play at each point a bandit of lowest "Gittins' index". This index can be computed separately for each bandit, by solving an optimal stopping problem.

The subtlety in the proof of Gittins' theorem is to give a tractable representation of the class of control policies available to the decision maker. In the original formulation (see, for example, [41, 75, 76, 39]), the class considered is feedback controls, as in a standard Markovian stochastic control problem; i.e. the system of bandits is modelled as a single Markov process, and the controls alter its transition probabilities. This formulation is restrictive, as it is not clear how it can be applied to a non-Markovian framework. Furthermore, as the control determines the filtration observed, it is also not clear how to introduce a general form of uncertainty aversion in this framework.

El Karoui and Karatzas [32] extend the argument of Gittins' theorem to the general case, without a Markovian assumption, by using Mandelbaum's [57] "allocation strategy" formulation of the class of control policies. In particular, they view the cost of the bandits as a fixed process. The effect of allocation is to delay the realization of these fixed costs, which results in a benefit to the decision maker due to the time-value of money.

In this paper, we will use a slight modification of Mandelbaum's allocation strategy to describe optimal strategies using a robust Gittins' theorem (without a Markovian assumption) via the classical argument given by Weber [75], but using El Karoui and Karatzas' [32] formulation.

Remark 2.1. In most of the literature on Gittins' theorem, maximization of rewards is usually considered. For convenience, as is common in the theory of nonlinear expectations, we will consider the minimization of costs instead. Our presentation of others' results is with the corresponding changes in sign.

Assumption 2.2. *Suppose that we have M bandits. The mth bandit is associated with a filtered probability space $\big(\Omega^{(m)}, \mathbb{P}^{(m)}, (\mathcal{F}_t^{(m)})_{t\geq 0}\big)$. Playing this bandit for the tth time realizes a non-negative bounded cost $h^{(m)}(t)$, where the process $(h^{(m)}(t))_{t\geq 1}$ is adapted to $(\mathcal{F}_t^{(m)})_{t\geq 0}$. We assume that $\mathcal{F}_0^{(m)} = \{\phi, \Omega^{(m)}\}$.*

The goal of the decision maker is to minimize the discounted total cost, for a given discount factor $\beta \in (0,1)$, when they can choose the order in which bandits are played.

Before considering a robust approach, we first outline the solution to this problem in a standard setting of classical expectation.

Definition 2.3. *The **Gittins index** at time $s \geq 0$ of the mth bandit is given by*

$$\gamma^{(m)}(s) = \operatorname*{ess\,inf}_{\tau \in \mathcal{T}^{(m)}(s)} \frac{\mathbb{E}^{\mathbb{P}^{(m)}}\big(\sum_{t=1}^{\tau} \beta^t h^{(m)}(s+t) \big| \mathcal{F}_s^{(m)}\big)}{\mathbb{E}^{\mathbb{P}^{(m)}}\big(\sum_{t=1}^{\tau} \beta^t \big| \mathcal{F}_s^{(m)}\big)}$$

where $\mathcal{T}^{(m)}(s)$ is the space of positive $(\mathcal{F}_{s+t}^{(m)})_{t\geq 0}$-stopping times[4] and the essential infimum is taken in $L^\infty(\mathcal{F}_s^{(m)})$.

Definition 2.4. *We define the* orthant probability space $\big(\bar{\Omega}, \bar{\mathbb{P}}, (\mathcal{F}(s))_{s\in\mathcal{S}}\big)$ *by*

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

where $\mathcal{S} := \mathbb{N}_0^M$. We write $\mathcal{F}(\infty) := \bigotimes_{m=1}^{\infty} \mathcal{F}_\infty^{(m)}$.

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

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

and

$$\mathfrak{T}(\mathcal{S}) := \Big\{S = (S^{(1)}, ..., S^{(M)}) \quad : \quad S^{(m)} \in \mathcal{T}^{(m)}\Big\}.$$

For $S \in \mathfrak{T}(\mathcal{S})$, we write $\mathcal{F}(S) := \bigotimes_{m=1}^{M} \mathcal{F}_{S^{(m)}}^{(m)}$ for $S \in \mathfrak{T}(\mathcal{S})$.

2.1.1 Classical Gittins theorem

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

Definition 2.5 (Mandelbaum [57]). *The **Mandelbaum allocation strategy** is an $\mathcal{S}$-valued random sequence $\big(\tilde{\eta}(n)\big)_{n\geq 0}$ such that*

(i) $\tilde{\eta}(0) = 0$

(ii) $\tilde{\eta}(n+1) = \tilde{\eta}(n) + e^{(m)}$ *for some $m \in \mathcal{M} =: \{1, ..., M\}$.*

[4] Equivalently, for $\tau \in \mathcal{T}^{(m)}(s)$, $s+\tau$ is an $(\mathcal{F}_t^{(m)})_{t\geq 0}$-stopping time.

(iii) $\{\tilde\eta(n+1)=\tilde\eta(n)+e^{(m)},\ \tilde\eta(n)=r\}\in\mathcal{F}(r)$ *for all* $m\in\mathcal{M}:=\{1,...,M\}$ *and for all* $r\in\mathcal{S}$.

Here, $e^{(m)}$ *denotes the* m*th unit vector in* $\mathcal{S}$*. We denote by* $\mathcal{A}$ *the family of all Mandelbaum allocation strategies.*

Theorem 2.6 (Gittins' theorem, as proved by El Karoui and Karatzas [32])**.** *Let* $\big(\tilde\eta^*(n)\big)_{n\geq 0}$ *be a Mandelbaum allocation strategy* $\big(\tilde\eta^*(n)\big)_{n\geq 0}$ *such that*

$$\gamma^{(m)}(\tilde\eta^{(m)}(n))=\min_{k\in\mathcal{M}}\gamma^{(k)}(\tilde\eta^{(k)}(n))\qquad \text{on each event}\quad \{\tilde\eta(n+1)=\tilde\eta(n)+e^{(m)}\}$$

for all $m\in\mathcal{M}$ *and* $n\geq 0$*. Then* $\big(\tilde\eta^*(n)\big)_{n\geq 0}$ *is an optimal solution to the optimization problem*

$$\inf_{\tilde\eta\in\mathcal{A}}\mathbb{E}^{\bar{\mathbb{P}}}\Big(\sum_{n=1}^{\infty}\sum_{m\in\mathcal{M}}\beta^n h^{(m)}(\tilde\eta^{(m)}(n))\big[\tilde\eta^{(m)}(n)-\tilde\eta^{(m)}(n-1)\big]\Big). \tag{2.1}$$

In particular, Theorem 2.6 says that the strategy which always plays the bandit with the minimum index minimizes the expectation of the total discounted cost.

Remark 2.7. 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 $\mathcal{M}$. In other words, we can define $(\rho_n)_{n\geq 0}$ such that $\{\rho_n=m\}=\{\tilde\eta(n+1)=\tilde\eta(n)+e^{(m)}\}$. We may then replace the objective equation (2.1) by

$$\inf_{\tilde\eta\in\mathcal{A}}\mathbb{E}^{\bar{\mathbb{P}}}\Big(\sum_{n=1}^{\infty}\beta^n h^{(\rho_{n-1})}(t_n^\rho)\Big)\quad\text{where}\quad t_n^\rho:=\sum_{k=0}^{n-1}\mathbb{I}(\rho_k=\rho_{n-1}). \tag{2.2}$$

Remark 2.8. Our paper considers the orthant filtration as the product of filtrations defined on different spaces. This is slightly different from Mandelbaum [57] (and thus El Karoui and Karatzas [32]) where the orthant filtration is considered as the join of filtrations defined on the same space. This technical difference will allow us to more easily define a 'Nonlinear expectation' which still carries some form of independence and 'time-consistency'. (See discussion in Section 3.2.)

Remark 2.9. In El Karoui and Karatzas [32], it is assumed that the cost process is predictable, instead of adapted, with respect to the filtration of the bandit. When using a classical expectation, there is no modelling difference between predictable and adapted cost processes (as one can just take the conditional expectation to reduce adapted costs to predictable costs). However, under a 'nonlinear expectation', this is not the case, so we give the more general result with adapted costs.

2.1.2 Robust Gittins index

Under a Markovian assumption, Caro and Gupta [20] consider a 'robust' Gittins index based on the Robust Bellman equation studied in Iyergar [46] and Nilim and El Ghaoui [58]. (Similar work is considered by Kim and Lim [51] with an additional penalty in the formulation.)

The following assumptions are used in Caro and Gupta [20] (translated into our notation):

(i) The cost process is driven by an underlying finite-state process $(X_t^{(m)})_{t\geq 0}$ taking values in $\mathcal{X}^{(m)}$, i.e. $h^{(m)}(t)=\tilde h^{(m)}(X_t^{(m)})$ for some deterministic function $\tilde h^{(m)}$.

(ii) Ambiguity is described by families of transition matrices, $(\mathcal{U}^{(m)})_{m\in\mathcal{M}}$ for the dynamics of $X^{(m)}$, which may vary in time.

The construction is then based on Whittle indexibility [77]. In particular, they reduce the problem to considering two bandits, where one bandit always generates a constant cost γ and the other bandit is identical to the mth bandit. The worst-case expected cost obtained when starting in state i in the mth bandit, $V^{(m)}(i)$, allowing any combination of transition rates, will then satisfy the robust dynamic programming principle, that is,

$$V^{(m)}(i) = \min\left(\tilde{h}^{(m)}(i) + \beta \sup_{P\in\mathcal{U}^{(m)}} \sum_{j\in\mathcal{X}^{(m)}} P_{ij}V^{(m)}(j), \quad \frac{\gamma}{1-\beta}\right), \qquad i \in \mathcal{X}^{(m)}. \tag{2.3}$$

Let $D^{(m)}(\gamma) \subseteq \mathcal{X}^{(m)}$ be the set of states for which it is optimal to rest the mth bandit when the reward of the constant bandit is γ. Caro and Gupta show that the robust bandit is Whittle indexible in the sense that $D^{(m)}(\gamma)$ increases monotonically from ϕ to $\mathcal{X}^{(m)}$ as γ increases from $-\infty$ to $+\infty$. The index of the mth bandit at state i is the unique value γ such that the player is indifferent between playing the mth bandit and the constant bandit.

This index can be characterized by

$$\tilde{\gamma}^{(m)}(i) = \inf_{\tau\in\mathcal{T}^{(m)},\tau\geq 1} \sup_{\mathbb{Q}\in\mathcal{Q}^{(m)}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t\tilde{h}^{(m)}(X_t^{(m)})\big|X_0^{(m)}=i\big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t\big|X_0^{(m)}=i\big)}. \tag{2.4}$$

where $\mathcal{Q}^{(m)}$ is the family of probabilities corresponding to the family of transitions $\mathcal{U}^{(m)}$.

Unfortunately, as discussed in Caro and Gupta [20], the robust Gittins index (2.4) does not yield a strategy optimizing the robust Bellman equation

$$V(i_1,...,i_K) = \min_{k\in[K]}\left(\tilde{h}^{(k)}(i_k) + \beta \sup_{P\in\mathcal{U}^{(k)}} \sum_{j\in\mathcal{X}^{(k)}} P_{i_kj}V(i_1,...,i_{k-1},j,i_{k+1},...,i_K)\right)$$

where $(i_1,...,i_K) \in \prod_{m=1}^M \mathcal{X}^{(m)}$.

In short, this non-optimality arises due to the fact that the robust Bellman equation introduces dependence between bandits. In particular, at equilibrium, the adversary (who determines the transition probabilities for each bandit) may choose differently depending on the state of *all* bandits, rather than just the bandit of interest.

The index (2.4) also can be interpreted as a Lagrangian relaxation of the optimal control problem (see also Gocgun and Ghate [43]). The natural question that arises is, 'Does this relaxation satisfy some adjusted notion of optimality?'

In this paper, we 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 through the martingale optimality principle, in order to address a control problem under an inconsistent nonlinear operator. We will show that the strategy given by the robust Gittins index satisfies this optimality criteria. We also allow the cost to be continuous valued and non-Markovian as in El Karoui and Karatzas [32]. This allows the study of various numerical methods to estimate our probabilistic state in the learning problem, whereas the numerical method in Caro and Gupta [20] is limited to finite state Markov process. A simple numerical example then allows us to observe some qualitative peculiarities given the interaction between uncertainty aversion and learning.

Remark 2.10. In a non-Markovian framework, Whittle indexibility is not well-defined. Hence, a reinterpretation of optimality is required to understand a solution to the multi-armed bandit problem under uncertainty aversion.

Remark 2.11. Li [56] considers a Bayesian formulation for the index but allowing for multiple priors. The focus is on describing how the set of uncertainty affects the index, but without proving any form of Whittle indexibility. Our models also verify and generalize these results.

3 Uncertainty, nonlinear expectation and optimality

In this section, we will outline how 'nonlinear expectation' operators can be used to model Knightian uncertainty. 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 build on the modelling framework of El Karoui and Karatzas [32] as proposed in Assumption 2.2.

We will first outline our setup and the additional assumptions we use in our study of the robust bandit problem. We will use a 'nonlinear expectation' $\mathcal{E}^{(m)}$ (Assumption 3.10) to model uncertainty on the space $\big(\Omega^{(m)}, \mathbb{P}^{(m)}, (\mathcal{F}_t^{(m)})_{t\geq 0}\big)$ of a single bandit, and then extend our uncertainty to the orthant joint space (Definition 2.4) via the combined nonlinear expectation $\mathfrak{E}$ (Definition 3.11). We will omit the superscript (m) when it is clear from context.

In order to avoid technical difficulties, we will make the following assumption on the cost processes.

Assumption 3.1. *For each $m \in \mathcal{M}$, there exists $C^{(m)} < \infty$ such that*

$$0 \leq h^{(m)}(t) \leq C^{(m)} \qquad \text{and} \qquad h^{(m)}(t) \to C^{(m)} \;\text{ as }\; t \to \infty \;\; \mathbb{P}\text{-a.s.}$$

Assumption 3.1 is purely technical. We may replace boundedness of $h^{(m)}$ by an integrability assumption on the total discounted cost (as in [32]); we then need to generalize the domain of the nonlinear expectation. We can also remove the assumption on the convergence of $h^{(m)}$ to its bound, but we then need to take more care to ensure that the stopping times we considered in (2.4) and elsewhere can be assumed to be a.s. finite. Given the discount factor, this assumption does not have a large impact on our modelling.

3.1 Nonlinear expectations and time consistency

Let us focus on the filtered probability space $\big(\Omega, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0}\big)$ modelling the returns from playing a single bandit. As in Peng [61], we define a nonlinear expectation as follows:

Definition 3.2. *A system of operators*

$$\mathcal{E}\left(\,\cdot\,|\mathcal{F}_t\right) : L^\infty(\mathbb{P}, \mathcal{F}_\infty) \to L^\infty(\mathbb{P}, \mathcal{F}_t) \quad ; \quad t \in \{0, 1, 2, ...\}$$

*is said to be an $(\mathcal{F}_t)_{t\geq 0}$-*consistent coherent nonlinear expectation *if it satisfies the following properties: for $X_n, X, Y \in L^\infty(\mathcal{F}_\infty)$ and $c \in L^\infty(\mathcal{F}_t)$, with all (in)equalities holding $\mathbb{P}$-a.s, we have*

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

(ii) $(\mathcal{F}_t)_{t\geq 0}$-Translation Equivariance*: $\mathcal{E}(X + c|\mathcal{F}_t) = \mathcal{E}(X|\mathcal{F}_t) + c$.*

(iii) Subadditivity*: $\mathcal{E}(X + Y|\mathcal{F}_t) \leq \mathcal{E}(X|\mathcal{F}_t) + \mathcal{E}(Y|\mathcal{F}_t)$.*

(iv) $(\mathcal{F}_t)_{t\geq 0}$-Positive Homogeneity*: If $c \geq 0$, then $\mathcal{E}(cX|\mathcal{F}_t) = c\,\mathcal{E}(X|\mathcal{F}_t)$.*

(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 $\mathcal{E}(X_n|\mathcal{F}_t) \to \mathcal{E}(X|\mathcal{F}_t)$ $\mathbb{P}$-a.s.*

(vi) $(\mathcal{F}_t)_{t\geq 0}$-consistency*: For $0 \leq s \leq t$, $\mathcal{E}(X|\mathcal{F}_s) = \mathcal{E}\big(\mathcal{E}(X|\mathcal{F}_t)\big|\mathcal{F}_s\big)$.*

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

Remark 3.3. For simplicity, we 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 [36, Corollary 4.38]). Moreover, if the operator $\mathcal{E}$ is induced by a BSDE (as in [27, 28, 61, 34] and many other papers), then the Lebesgue property typically follows from the L^2-continuous dependence of the BSDE on its terminal value.

Remark 3.4. It is also known (see e.g. Detlefsen and Scandolo [30]) that any coherent nonlinear expectation satisfies the $(\mathcal{F}_t)$-*regularity* property. That is, for any $X, Y \in L^\infty(\mathcal{F}_\infty)$ and $A \in \mathcal{F}_t$,

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

In particular, $\mathcal{E}\big(X\mathbb{I}_A\big|\mathcal{F}_t\big) = \mathbb{I}_A\mathcal{E}\big(X\big|\mathcal{F}_t\big)$.

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

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

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

Proposition 3.6. *The operator $\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_\tau\big)$ satisfies the conditions of Definition 3.2 with s and t replaced by stopping times.*

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 (over a finite horizon) given by Artzner et al. [5], see also Föllmer and Schied [36] and Frittelli and Rosazza-Gianin [37]. Extensions to a dynamic setting are also considered by Detlefsen and Scandolo [30], Föllmer and Schied [36] and Riedel [67]. We state a version of this result which is dynamic over stopping times.

Theorem 3.7. *Let $\mathcal{E}$ be a consistent coherent nonlinear expectation. If there exists $T < \infty$ such that $\mathcal{F}_\infty = \mathcal{F}_T$, 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, $\mathcal{Q} \subseteq \{\mathbb{Q}$ probability measures $: \mathbb{Q} \approx \mathbb{P}\}$, and the essential supremum is taken in $L^\infty(\mathcal{F}_\tau, \mathbb{P})$.

Proof. See Föllmer and Schied [36, Theorem 11.22] (with further discussion in Föllmer and Penner [35]). 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 3.8. Theorem 3.7 can be obtained by construction via considering the stability of the pasting in the family $\mathcal{Q}$. (See e.g. Bion-Nadal [15] and Artzner et al. [6]).

3.2 Uncertainty on multiple bandits

In the classical Gittins theorem, independence is crucial to separate the behaviour of different bandits. In the robust representation (Theorem 3.7) 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 bandits becomes ambiguous, as statistical independence is based on the probability measure. Thanks to our explicit construction of the space (Definition 2.4), we can explicitly construct a nonlinear expectation space where each bandit remains independent.

Remark 3.9. In [63], 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 bandits, 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 [63] in this setting.

Let us make the last universal assumption in our paper, which describes model uncertainty for each individual bandit in our problem, inspired by the robust representation (Theorem 3.7).

Assumption 3.10. *For each $m \in \mathcal{M}$, we have a $(\mathcal{F}_t^{(m)})_{t\geq 0}$-consistent coherent nonlinear expectation, $(\mathcal{E}^{(m)}(\,\cdot\,|\mathcal{F}_\tau))_{\tau\in\mathcal{T}^{(m)}}$ defined on the space $L^\infty(\Omega^{(m)}, \mathbb{P}^{(m)})$ which admits the representation*

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

whenever $S^{(m)}$ is an $(\mathcal{F}_t^{(m)})$-stopping time.

Definition 3.11. *We define the* partially consistent orthant nonlinear expectation $(\mathfrak{E}_S)_{S\in\mathfrak{T}(\mathcal{S})}$*, to be the family of operators*

$$\begin{aligned} \mathfrak{E}_S : L^\infty\big(\bar\Omega, \bar{\mathbb{P}}, \mathcal{F}(\infty)\big) \quad &\longrightarrow \quad L^\infty\big(\bar\Omega, \bar{\mathbb{P}}, \mathcal{F}(S)\big) \\ X \quad &\longmapsto \quad \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\big(X\big|\mathcal{F}(S)\big) \end{aligned}$$

where $\mathcal{Q} := \big\{ \bigotimes_{m=1}^M \mathbb{Q}^{(m)} \;\text{ for }\; \mathbb{Q}^{(m)} \in \mathcal{Q}^{(m)}\big\}$, with $\mathcal{Q}^{(m)}$ as in Assumption 3.10.

We also write $\mathfrak{E}(\cdot)$ for $\mathfrak{E}_0(\cdot)$.

Remark 3.12. As $\bar{\mathbb{P}} = \bigotimes_{m=1}^M \mathbb{P}^{(m)}$ 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 3.13. *The system of operators $(\mathfrak{E}_S)$ satisfies the following properties.*

(i) The properties (i)-(v) in Definition 3.2 (with appropriate replacements on the operator and σ-algebra) 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 hold for $\mathfrak{E}$).

(ii) Sub-consistency: For $S, S' \in \mathfrak{T}(\mathcal{S})$ with $S \leq S'$, we have

$$\mathfrak{E}_S\,(\,\cdot\,) \leq \mathfrak{E}_S\,(\mathfrak{E}_{S'}\,(\,\cdot\,)) \quad \bar{\mathbb{P}}\text{-a.s.}$$

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 $(\bar\Omega, \mathcal{F}(\infty))$ given by

$$Y(\omega^{(1)}, ..., \omega^{(M)}) = X^{(1)}(\omega^{(1)}) \times \cdots \times X^{(M)}(\omega^{(M)}) \qquad \bar{\mathbb{P}}\text{-a.s.},$$

where, for each $m \in \mathcal{M}$, we have a non-negative random variable $X^{(m)}$ defined on $(\Omega^{(m)}, \mathcal{F}^{(m)}_\infty)$. Then

$$\mathfrak{E}_S(Y) = \mathcal{E}^{(1)}\left(X^{(1)}\Big|\mathcal{F}^{(1)}_{S^{(1)}}\right) \times \cdots \times \mathcal{E}^{(M)}\left(X^{(M)}\Big|\mathcal{F}^{(M)}_{S^{(M)}}\right) \quad \bar{\mathbb{P}}\text{-a.s.}$$

(iv) Marginal projection: For a given $m \in \mathcal{M}$, let X be a random variable defined on $(\Omega^{(m)}, \mathcal{F}^{(m)}_\infty)$. Define $\tilde X : \bar\Omega \to \mathbb{R}$ by $\tilde X(\omega^{(1)}, ..., \omega^{(M)}) = X(\omega^{(m)})$. We then have

$$\mathfrak{E}_S(\tilde X) = \mathcal{E}^{(m)}\big(X\big|\mathcal{F}^{(m)}_{S^{(m)}}\big) \quad \bar{\mathbb{P}}\text{-a.s.}$$

Proof. See Proposition C.4 in the appendix. □

Remark 3.14. We deliberately choose our nonlinear expectation $\mathfrak{E}$ to be defined on a product space to simplify our discussion on the existence of the operator. In fact, one can simply weaken our assumption by having a nonlinear expectation $\mathfrak{E}$ on a joint filtration (as in El Karoui and Karatzas [32]) such that the Proposition 3.13 holds. All proofs are identical, except the proof of Theorem B.6 (in the Appendix), where we need an extra step to show that the product of the marginal probability measure is also a probability measure considered under the robust representation of $\mathfrak{E}$.

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 3.2, 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 3.15. 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

$$\sup_{\mathbb{Q},\tilde{\mathbb{Q}}} \mathbb{E}^{\mathbb{Q}\otimes\tilde{\mathbb{Q}}}\big(f(X,\tilde{X})\big) = \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\sup_{\tilde{\mathbb{Q}}} \mathbb{E}^{\tilde{\mathbb{Q}}}\big(f(x,\tilde{X})\big)\Big|_{x=X}\Big) = \max\Big(\frac{f(0,0)+f(0,1)}{2}, \frac{f(1,0)+f(1,1)}{2}\Big)$$

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 a nonlinear expectation using the supremum over the families $\mathcal{Q}$ 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

$$\mathfrak{E}\Big(\mathfrak{E}_{(0,1)}\Big(f(X,\tilde{X})\Big)\Big) \neq \mathfrak{E}\Big(\mathfrak{E}_{(1,0)}\Big(f(X,\tilde{X})\Big)\Big).$$

3.3 Optimality

We have discussed in the previous section that the robust Gittins index (2.4) in the sense of Caro and Gupta [20] is not optimal, as it does not lead to a solution of the robust Bellman equation (discussed in [46, 58]). In order to understand what sense of optimality the robust index strategy *does* satisfy, we will first consider a type of optimality criteria used by El Karoui and Karatzas [32].

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 $\big(g^\rho(n)\big)_{n\geq 1}$. 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 (2.2). We can also define the filtration of information obtained up to time n when following ρ by

$$\mathcal{G}_n^\rho := \Big\{A \in \mathcal{F}(T) \ : \ A \cap \{\tilde{\eta}(n) = r\} \in \mathcal{F}(r) \quad \forall r \in \mathcal{S}\Big\}, \tag{3.1}$$

where $\tilde{\eta}$ is the corresponding Mandelbaum allocation sequence (Definition 2.5, Remark 2.7). We will discuss this filtration in detail in Remarks 4.13 and 4.14.

Remark 3.16. 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 (2.2)) is also adapted with respect to $\mathcal{G}_n^\rho$.

Suppose that we are given a nonlinear expectation operator $\mathfrak{E}$, as in Definition 3.11, and consider a minimization problem over the space of Mandelbaum allocation strategies, as represented by their equivalent form ρ (Remark 2.7). The process ρ not only describes our strategy and the corresponding cost, but also determines the observed filtration. Therefore, at any point in time, it does not make sense to compare strategies unless those strategies yield the same information at the considered time.

Definition 3.17. *We say strategies ρ and ρ' are historically equivalent at time N, denoted by $\rho \sim_N \rho'$, if $\rho_n = \rho'_n$ for all $n \leq N$.*

Remark 3.18. For every strategy ρ, we have $\rho \sim_0 \rho^*$.

We can now give a standard form of optimality which is often considered when we have a consistent nonlinear expectation operator.

Definition 3.19. *We say a strategy ρ^* is a* strong optimum *if, for every $N \geq 0$, for every strategy ρ such that $\rho \sim_N \rho^*$, we have*

$$\mathfrak{E}\bigg(\mathbb{I}_A\bigg(\sum_{n=N+1}^{\infty} g^{\rho^*}(n)\bigg)\bigg) \leq \mathfrak{E}\bigg(\mathbb{I}_A\bigg(\sum_{n=N+1}^{\infty} g^{\rho}(n)\bigg)\bigg) \qquad \textit{for all } A \in \mathcal{G}_N^{\rho} (= \mathcal{G}_N^{\rho^*}).$$

Remark 3.20. When $\mathfrak{E}$ is replaced by an $(\mathcal{F}_n)$-consistent nonlinear expectation and $(\mathcal{G}_N^{\rho})$ is replaced by $(\mathcal{F}_N)$, strong optimality simplifies to

$$\mathcal{E}\bigg(\sum_{n=N+1}^{\infty} g^{\rho^*}(n)\bigg|\mathcal{F}_N\bigg) = \operatorname*{ess\,inf}_{\rho\sim_N\rho^*} \mathcal{E}\bigg(\sum_{n=N+1}^{\infty} g^{\rho}(n)\bigg|\mathcal{F}_N\bigg).$$

A standard approach to tackle decision making under a time-inconsistent (nonlinear expectation) operator is to define 'the optimal strategy' through the solution of the robust Bellman equation [46, 58], as considered in Caro and Gupta [20]. 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.

3.4 C-Optimality

In the bandit setting, our nonlinear expectation is not necessarily (time-)consistent. In order to understand the Gittins index strategy under an inconsistent operator, we propose an alternative notion of optimality, which is inspired by martingale optimality.

For motivation, consider an $(\mathcal{F}_n)$-consistent nonlinear expectation $\mathcal{E}$. Suppose that we wish to solve the minimization problem

$$V_N = \operatorname*{ess\,inf}_{\rho} \mathcal{E}\bigg(\sum_{n=N+1}^{\infty} g^{\rho}(n)\bigg|\mathcal{F}_N\bigg).$$

For a given strategy ρ, we define a process $X_N^{\rho} := \sum_{n=1}^{N} g^{\rho}(n) + V_N$. Under mild conditions, we know from the martingale optimality principle that (X_N^{ρ}) is an $\mathcal{E}$-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. [26, Theorem 8]), we can write

$$X_N^{\rho} := M_N^{\rho} + \sum_{n=1}^{N} C^{\rho}(n)$$

where (M_N^{ρ}) is an $\mathcal{E}$-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, for every ρ,

$$\mathcal{E}\bigg(\sum_{n=N+1}^{\infty}\Big(g^{\rho}(n)-C^{\rho}(n)\Big)\bigg|\mathcal{F}_N\bigg)=V_N.$$

Moreover, for an optimal strategy ρ^*, $\sum_{n=N+1}^{L} C^{\rho^*}(n) = 0$ and for any strategy ρ, $\sum_{n=N+1}^{L} C^{\rho}(n) \geq 0$. This implies that

$$\sum_{n=N+1}^{L} C^{\rho^*}(n) \leq \sum_{n=N+1}^{L} C^{\rho}(n) \qquad \text{for all } N \text{ and } L \geq N+1.$$

Inspired by the analysis above, we propose an alternative notion of optimality in an inconsistent setting.

Definition 3.21. *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 $(C_N^{\rho}(n))_{N,n\geq N+1,\rho\sim_N\rho^*}$ (called a (sub-)compensator) such that*

(i) $n \mapsto C_N^{\rho}(n)$ *is a* $(\mathcal{G}_n^{\rho})$*-predictable process,*

(ii) $N \mapsto C_N^{\rho}(n)$ *is non-increasing,*

(iii) For every strategy $\rho \sim_N \rho^*$,

$$\mathfrak{E}\bigg(\mathbb{I}_A\sum_{n=N+1}^{\infty}\Big(g^{\rho}(n)-C_N^{\rho}(n)\Big)\bigg) \geq \mathfrak{E}\Big(\mathbb{I}_A V_N\Big) \qquad \textit{for all} \quad A \in \mathcal{G}_N^{\rho} \tag{3.2}$$

with equality for $\rho = \rho^*$,

(iv) For every strategy $\rho \sim_N \rho^*$,

$$\sum_{n=N+1}^{L} C_N^{\rho^*}(n) \leq \sum_{n=N+1}^{L} C_N^{\rho}(n) \qquad \textit{for all} \quad L \geq N+1. \tag{3.3}$$

We can see $\big(C_N^{\rho}(n)\big)$ acts as '(sub-)compensator' to the cost, and V_N acts as the value function. This approach is loosely related to the capital requirement approach discussed by Frittelli and Scandolo [38]. We can interpret Definition 3.21 as requiring that the (sub-)compensators $(C_N^{\rho}(n))$

(i) are known one-step in advance before observing the cost (i).

(ii) consistently (sub-)compensate the cost. In particular, as time elapses, we obtain more information and thus require the same amount, or possibly less, to (sub-)compensate (ii).

(iii) complement the extra cost occurred for a sub-optimal strategy (iii).

(iv) are bounded below by the compensator of a particular strategy ρ^*, which we call 'optimal' (iv).

Remark 3.22. We have mentioned the robust Bellman equation [46, 58] 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 [73] and Pollak [64] and developed further in Peleg and Yaari [60] and Koopmans [53]. Recent extensions include Björk and Murgoci [17], Björk, Khapko and Murgoci [16], Yong [79] and Hu, Jin and Zhou [44]. The essential idea of this approach is to use backward induction at the level of controls. That is, 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 $\big\{\rho \ : \ (\rho_{n+1},...,\rho_L) = (\rho^*_{n+1},...,\rho^*_L)\big\}$. This idea is then extended by searching for (subgame perfect) Nash equilibria, to allow for non-uniqueness of the optimal controls.

We have discussed in Section 2.1.2 that the robust Bellman approach may introduce some dependence between bandits in our system. Hence, the Gittins index strategy is not optimal under that approach. On the other hand, when considering a system of bandits, the future state is 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.

The notion of C-optimality could be loosely interpreted as the mixture of the robust Bellman approach and the Strotz–Pollak approach. In particular, we can interpret the compensated process as propagating the value function backward in time, as in the robust Bellman approach, but only the one-step value is propagated backward in time, which is related to the Strotz–Pollak approach.

3.5 Endowment effect

One natural question to ask is whether we can give an interpretation of C-optimality (Definition 3.21) in terms of classical strong optimality (Definition 3.19). To see this, we will consider how an endowment affects strong optimality.

Example 3.23. 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

$$\mathfrak{E}(H) < \mathfrak{E}(G) \quad \text{but} \quad \mathfrak{E}\Big(H-\frac{H+G}{2}\Big) > \mathfrak{E}\Big(G-\frac{H+G}{2}\Big). \tag{3.4}$$

From these inequalities, we see that, without any endowment, we strictly prefer H to G, whereas our preference reverses with an endowment $(H+G)/2$. In the classical linear expectation theory (where the classical Gittins theorem holds), we know that an endowment does not affect our preferences over the strategy.

In this section, we will show that C-optimality is nearly equivalent to 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 (in particular, Definition 3.19(iii)-(iv)).

Proposition 3.24. *Let ρ^* be a C-optimal strategy with a predictable compensator $(C^{\rho^*}_N(n))$. Then for every $\rho\sim_N\rho^*$ and $A\in\mathcal{G}^{\rho^*}_N$,*

$$\mathfrak{E}\bigg(\mathbb{I}_A\sum_{n=N+1}^{\infty}\big(g^{\rho^*}(n)-C^{\rho^*}_N(n)\big)\bigg) \leq \mathfrak{E}\bigg(\mathbb{I}_A\sum_{n=N+1}^{\infty}\big(g^{\rho}(n)-C^{\rho^*}_N(n)\big)\bigg). \tag{3.5}$$

Let consider the case when $\mathcal{G}^{\rho}_N=\mathcal{F}_N$ for every strategy ρ and pretend that $\mathfrak{E}$ is an $(\mathcal{F}_n)$-consistent nonlinear expectation operator. Then (3.5) says that C-optimality implies strong optimality, when our agent is given the predictable endowment $-\sum_{n=N+1}^{\infty}C^{\rho^*}_N(n)$

at time N. We will now show that a converse result also holds, when our operator is consistent.

Definition 3.25. *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 $A \in \mathcal{G}_N^\rho$,

$$\mathcal{E}\bigg(\mathbb{I}_A \sum_{n=N+1}^{\infty} \Big(g^{\rho^*}(n) - D_N(n)\Big)\bigg) \leq \mathcal{E}\bigg(\mathbb{I}_A \sum_{n=N+1}^{\infty} \Big(g^{\rho^*}(n) - D_N(n)\Big)\bigg) \tag{3.6}$$

or equivalently,

$$\mathcal{E}\bigg(\sum_{n=N+1}^{\infty} \Big(g^{\rho^*}(n) - D_N(n)\Big)\bigg|\mathcal{F}_N\bigg) \leq \mathcal{E}\bigg(\sum_{n=N+1}^{\infty} \Big(g^{\rho^*}(n) - D_N(n)\Big)\bigg|\mathcal{F}_N\bigg).$$

Proposition 3.26. *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 $V_N = \mathcal{E}\big(\sum_{n=N+1}^{\infty}\big(g^{\rho^*}(n) - D_N(n)\big)\big|\mathcal{F}_N\big)$. □

In the coming section, we will show that Gittins theorem holds in the sense of guaranteeing C-optimality under an operator $\mathfrak{E}$. This means that we prove that Gittins theorem yields a (strong) optimum up to some predictable endowment.

Remark 3.27. It is an open question under which conditions the C-optimum is unique. In the most trivial case when our operator $\mathfrak{E}$ is simply a classical expectation, the endowment never affects our evaluation; thus it is reduced to the uniqueness of the value function in the classical setting.

4 Overview of bandits under uncertainty

Let us recall that the objective of our problem is to dynamically allocate a single resource amongst M bandits to minimize the total discounted cost. We have made a few assumptions to model uncertainty in the cost process, which can be found in Assumptions 2.2, 3.1 and 3.10.

We have also introduced Mandelbaum allocation strategies (Definition 2.5) and the equivalent notation ρ (Remark 2.7) representing the choice of our control. We are now ready to establish a robust Gittins theorem with optimality in the sense of Definition 3.21. Our robust Gittins theorem generalizes the result of El Karoui and Karatzas [32] to the uncertain case. One may also see this result as providing a sense of optimality for the index strategy considered by Caro and Gupta [20] and Li [56].

4.1 Robust Gittins theorem

We will first give an alternative definition of the robust Gittins index inspired by Weber [75], which is more convenient to use in our analysis.

Definition 4.1. *For each $s \geq 0$, we define the* robust Gittins index *of the mth bandit by*

$$\gamma^{(m)}(s) := \operatorname*{ess\,inf}\left\{\gamma : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}^{(m)}(s)} \mathcal{E}^{(m)}\left(\sum_{t=1}^{\tau}\beta^t\big(h^{(m)}(s+t)-\gamma\big)\;\middle|\;\mathcal{F}_s^{(m)}\right) \leq 0\right\} \tag{4.1}$$

where $\mathcal{T}^{(m)}(s)$ is the space of positive $(\mathcal{F}^{(m)}_{s+t})_{t\geq 0}$-stopping times[5] *and the outer essential infimum is taken in $L^\infty(\mathcal{F}^{(m)}_s)$.*

By using the results proved in the later sections, we can write the robust Gittins index explicitly. We present this result here for clarity, but make no use of it in subsequent arguments.

Theorem 4.2. *Let $\gamma(s)$ be the robust Gittins index (Definition 4.1) (with superscript (m) omitted). Then*

$$\gamma(s) = \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\big)}$$

where $\mathcal{Q}$ is the family of probability measures defined in Theorem 3.7.

Proof. See Theorem C.3 in the appendix. □

Recall that $\mathfrak{E}$ is the partially consistent orthant nonlinear expectation induced by the family $\big(\mathcal{E}^{(m)}\big)_{m\in\mathcal{M}}$ as given in Definition 3.11. We can obtain an optimal allocation strategy by considering the following theorem.

Theorem 4.3 (Robust Gittins theorem)**.** *Suppose that for each $m\in\mathcal{M}$, $(\mathcal{F}^{(m)}_t)_{t\geq 0}$ is generated by some underlying process $(\xi^{(m)}_t)_{t\geq 1}$. Let $\psi^{(m)}_n$ be the total number of trials of the mth bandit before the nth play of the system. i.e. $\psi^{(m)}_n := \sum_{k=0}^{n-1}\mathbb{I}(\rho^*_k = m)$ (given an allocation strategy ρ^* up to time $n-1$).*

Then the allocation strategy ρ^* given (recursively) by

$$\rho^*_n := \min\Big\{m\in\mathcal{M} \quad:\quad m\in\arg\min_k \gamma^{(k)}(\psi^{(k)}_n)\Big\}$$

is C-optimal (Definition 3.21) under $\mathfrak{E}$ for the cost

$$g^\rho(n) = \beta^n h^{(\rho_{n-1})}(t^\rho_n) \quad \textit{where} \quad t^\rho_n = \sum_{k=0}^{n-1}\mathbb{I}(\rho_k = \rho_{n-1}).$$

Remark 4.4. We choose ρ^* to be the minimum value in the (random) set of minimum Gittins index machines $\{\arg\min_k \gamma^{(k)}(\psi^{(k)}_n)\}$ as a simple method of symmetry breaking, in order to avoid complexities due to measurable selection. In fact, any choice of $\rho^*_n \in \{\arg\min_k \gamma^{(k)}(\psi^{(k)}_n)\}$ also yields C-optimality.

Remark 4.5. The robust Gittins theorem states that an optimal choice is given by always playing a bandit with the lowest robust Gittins index. At each time, the indices of unplayed bandits do not change. This leads to a form of consistency in the values associated with different bandits, even though $\mathfrak{E}$ is not consistent.

4.2 Sketch of the Proof

We will separate the proof into two parts: In Part A, we analyze a one-armed bandit in a robust setting. In Part B, we combine M bandits together. The main body of the rigorous proof can be found in Appendices A and B (respectively) as self-explained sections. We summarize the structure and approach of the proof here.

4.2.1 One-armed bandit optimality

We begin by considering play of the mth machine (with the superscript (m) omitted).

[5] Equivalently, for $\tau\in\mathcal{T}(s)$, $s+\tau$ is an $(\mathcal{F}_t)_{t\geq 0}$-stopping time.

Step A.1 Observe that the robust Gittins index is the *minimum* compensation for which we are willing to continue to play the bandit (with compensation).

By minimality, the net expected cost under optimal play must be zero (Theorem A.3), 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, for any subsequent stopping time $\tau \in \mathcal{T}(s)$, we have

$$\mathcal{E}\left(\sum_{t=1}^{\tau} \beta^t \left(h(s+t) - \gamma(s)\right) \,\middle|\, \mathcal{F}_s \right) \geq 0. \tag{4.2}$$

Step A.2 We view the process γ as the 'average' cost of playing the bandit. 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

$$\sigma(s,\gamma(s)) := \inf\{\theta \geq 1 : \gamma(s+\theta) > \gamma(s)\} \tag{4.3}$$

yields equality in (4.2) (Theorem A.10).

Step A.3 Imagine that, whenever the bandit (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 A.14). The offered compensation can be written as a running maximum of the robust Gittins index process and we can express the expected return

$$\mathcal{E}\left(\sum_{t=1}^{\infty} \beta^t \big(h(t) - \Gamma(t)\big) \right) = 0 \qquad \text{where} \quad \Gamma(t) := \max_{0\leq\theta\leq t-1} \gamma(\theta). \tag{4.4}$$

With the compensation 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 A.16), i.e.

$$\mathcal{E}\left(\sum_{t=N+1}^{\infty} \beta^t \left(h(t) - \Gamma(t)\right) \middle| \mathcal{F}_N \right) \leq 0 \quad \text{for all} \quad N = 0, 1, \ldots \tag{4.5}$$

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 (4.5)) due to the discount effect of the delay.

By (4.4), the total reward of this game is zero. Therefore, the delay of getting the reward must result in a possibly worse outcome. In Theorem A.18, we use this observation, together with the robust representation (Assumption 3.10) 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 $[0, 1]$,

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\infty} \alpha(t)\beta^t \big(h(t) - \Gamma(t)\big) \right) \geq -\epsilon. \tag{4.6}$$

Remark 4.6. 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 bandit. In step A.3, we choose the compensator such that the total expected return is zero but the bandit 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 $\mathcal{E}(\alpha X|\mathcal{F}_t) \geq \mathcal{E}(X|\mathcal{F}_t)$.

4.2.2 Information structures for multi-armed bandits

We now consider combining play over multiple machines.

To retain consistency for a single bandit, 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 bandits, which we do via an orthogonal product space. This restriction does not allow us to directly implement Mandelbaum's [57] original approach for a dynamic allocation strategy (Definition 2.5). 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 $\Omega^{(m)}$ where our single-bandit nonlinear expectation is defined.

The importance of decomposing a strategy on the multi-armed bandit to strategies for one-armed bandits can be seen in the proof of El Karoui and Karatzas [32, Equation 5.1] (via Whittle's approach [76]), and is described more explicitly in their continuous time paper [33, Equation 6.9].

In order to overcome this difficulty, we introduce a class of allocation strategies where there is a component associated to the stopping times of the marginal filtrations. This component allows us to connect and separate the space of multiple bandits to the marginal space of each single bandit.

Our class of allocation strategies consists of two components (τ, p). The collection of random times $\tau = (\tau_k^{(m)})_{k\geq 0, m\in\{1,\dots,M\}}$ will identify the duration for which will play the mth bandit, the kth time we start to play. This sequence is chosen based on historical observations of the mth bandit only, that is, the random times $\sum_{k=0}^{K} \tau_k^{(m)}$ are $(\mathcal{F}_t^{(m)})_{t\geq 0}$-stopping times for all $K \geq 0$. Once we play a bandit for $\tau_k^{(m)}$ trials, we will then reconsider which bandit to play. Our choice of new bandit (which may be the same as before) will be described by the sequence (p_n) taking values in $\{1,\dots,M\}$, and may depend on information from all bandits. The allocation strategy can be defined formally as follows:

Definition 4.7. *We say* $\tau := \left(\tau_k^{(m)}\right)_{k\geq 0, m\in\mathcal{M}}$ *is a* family of time allocation sequences *if*

(i) For each m, $(\tau_k^{(m)})_{k\geq 0}$ *is a sequence of non-negative random times defined on the space* $(\Omega^{(m)}, \mathcal{F}_\infty^{(m)})$.

(ii) $\sum_{i=0}^{k} \tau_i^{(m)}$ *is an* $(\mathcal{F}_t^{(m)})_{t\geq 0}$*-stopping time for all* $k \geq 0$.

Intuitively, the random sequence (p_n) is allowed to depend on all prior observations from all bandits. For the sake of precise bookkeeping we need to record, at each moment, how many times we have already played each bandit. This leads to the following definition.

Definition 4.8. *Given a family of time allocation sequences* τ*, we say a sequence of random variables* $(\eta_n)_{n\geq 0}$ *taking values in* $\mathcal{S} = \mathbb{N}_0^M$ *is a* recording sequence *associated to* τ*, with corresponding* choice sequence $(p_n)_{n\in\mathbb{N}_0}$ *taking values in* $\mathcal{M}$*, if*

(i) $\eta_0 = (0, \dots, 0)$*, and*

(ii) $\eta_{n+1} = \eta_n + e^{(p_n)}$.

The choice process p_n *satisfies*

(iii) for all $k \in \mathcal{M}$ *and* $r \in \mathcal{S}$,

$$\{p_n = k\} \cap \{\eta_n = r\} \in \mathcal{F}(\Psi_r) = \bigotimes_{m=1}^{M} \mathcal{F}^{(m)}_{\Psi_r^{(m)}}$$

where $\Psi_r^{(m)} := \sum_{i=0}^{r^{(m)}-1} \tau_i^{(m)}$. *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

$$\mathcal{G}_n^{(\tau,p)} := \Big\{A \in \mathcal{F}(T) \ : \ A \cap \{\eta_n = r\} \in \mathcal{F}(\Psi_r) \quad \forall r \in \mathcal{S}\Big\} \tag{4.7}$$

where $\Psi_r^{(m)} = \sum_{i=0}^{r^{(m)}-1} \tau_i^{(m)}$.

Remark 4.9. We can see in Definition 4.8(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 4.10. *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 4.11. Suppose there are two bandits. The first bandit gives only 2 outcomes: $\{w, l\}$. Consider the strategy of playing the first bandit until we see the first l. Then we swap to the second bandit for two trials and swap back to the first bandit and repeat the same procedure.

In this case, we define $(X_t)_{t\geq 1}$ to be the outcome of the first bandit 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 2.7, we can express a Mandelbaum allocation strategy (Definition 2.5) 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 4.12. *Define the random variable ρ_n to be the bandit 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 C.8 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 4.13. Observe that if ρ is the simple form of (τ, p) and we denote the time allocation sequence $\imath = \big(1, 1, 1, ...\big)$, then $(\imath, \rho)$ is an allocation strategy which yields the same decisions as (τ, p). In particular, we have $(\imath, \rho) \sim (\tau, p)$.

Furthermore, one can check that the recording sequence corresponding to $(\imath, \rho)$ is exactly the Mandelbaum allocation strategy (Definition 2.5). In particular, we can explicitly construct a one-to-one correspondence between our equivalence class of admissible strategies (Definition 4.10) and Mandelbaum allocation strategies, and we have $\mathcal{G}_n^{(\imath,\rho)} = \mathcal{G}_n^{\rho}$ in (3.1).

Remark 4.14. Assume that, for $m \in \mathcal{M}$, the filtration $(\mathcal{F}_t^{(m)})$ is generated by an underlying real process $(\xi_t^{(m)})_{t\geq 1}$ defined on the space $(\Omega^{(m)}, \mathcal{F}_\infty^{(m)})$, i.e.

$$\mathcal{F}_0^{(m)} = \big\{\phi, \Omega^{(m)}\big\} \quad \text{and} \quad \mathcal{F}_t^{(m)} = \sigma\big(\xi_1^{(m)}, \xi_2^{(m)}, ..., \xi_t^{(m)}\big).$$

If we parameterize our actions by a simple form strategy $(\imath, \rho)$ 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

$$\xi_n^{\rho} := \xi^{(\rho_{n-1})}_{\eta_n^{(\rho_{n-1})}} = \sum_{m=1}^{M} \sum_{t=1}^{\infty} \xi_t^{(m)} \mathbb{I}(\rho_{n-1} = m,\ \eta_n^{(m)} = t).$$

We define the *observed filtration* by $\mathcal{H}_0^{\rho} = \{\phi, \bar{\Omega}\}$ and $\mathcal{H}_n^{\rho} := \sigma\,(\xi_1^{\rho}, ..., \xi_n^{\rho})$. We prove, in the appendix, that the observed filtration agrees with that used in Definition 4.8 when considering measurability of ρ. That is

$$\mathcal{H}_n^{\rho} = \mathcal{G}_n^{(\imath,\rho)} = \{A \in \mathcal{F}(T) \ \ : \ \ A \cap \{\eta_n = r\} \in \mathcal{F}(r)\ \} \tag{4.8}$$

where η is the recording sequence corresponding to $(\imath, \rho)$.

4.2.3 Multi-armed bandit optimality

We can now give the second half of the proof for the robust Gittins index theorem where we will consider 0 as our reference value function.

In order to prove the optimality of the robust Gittins' strategy, we define the target function for an allocation strategy by

$$V(\tau, p) := \mathfrak{E}\bigg(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1})}(t_n^{\rho}) - \Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\bigg) \ : \ t_n^{\rho} := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1}) \tag{4.9}$$

where ρ is the simple form derived from (τ, p) and $\big(\Gamma^{(m)}(t)\big)$ is the running max of the robust Gittins index of the mth bandit, as considered in (4.4).

Step B.1 Suppose that we have M bandits, with associated indifference rewards $(\Gamma^{(m)})_{m\in\mathcal{M}}$ as in step A.3. If we mix the play of these bandits, this is equivalent to taking a break in a single bandit to play the others. This delay will result in a possibly worse outcome (Equation (4.6) in step A.4).

In Theorem B.6, 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 $\bigotimes_{m=1}^{M} \mathbb{Q}^{(m)} \in \mathcal{Q}$ such that

$$\begin{aligned} V(\tau,p) &\geq \mathbb{E}^{\bigotimes_{m=1}^{M}\mathbb{Q}^{(m)}}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho}) - \Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right) \\ &= \sum_{m=1}^{M}\mathbb{E}^{\mathbb{Q}^{(m)}}\left(\sum_{t=1}^{\infty}\tilde{\alpha}^{(m)}(t)\beta^t\left(h^{(m)}(t) - \Gamma^{(m)}(t)\right)\right) \quad \geq \quad -M\epsilon \end{aligned}$$

where $\tilde{\alpha}^{(m)}(t)$ is the delay effect on the mth bandit due to playing other bandits.

As ϵ is arbitrary, it follows that, for all allocation strategies (τ, p),

$$V(\tau,p) \geq 0. \tag{4.10}$$

Step B.2 In step A.2, we noticed that the total expected loss of a single bandit between S and S' is zero, for S and S' the consecutive stopping times when the robust Gittins index hits a new maximum (Equation (4.3)). We use this fact to construct a family of time allocation sequences as a candidate optimal strategy.

Define (inductively) $S_k^{(m)} := \sum_{l=0}^{k-1}\sigma_l^{(m)}$ and

$$\sigma_k^{(m)} := \inf\left\{\theta \geq 1 : \gamma^{(m)}(S_k^{(m)} + \theta) > \gamma^{(m)}(S_k^{(m)})\right\}. \tag{4.11}$$

Using our construction of 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 result from step A.2, that $\sigma_k^{(m)}$ yields equality in (4.2), to show that, for any p, with the choice of time allocation sequences $\sigma = (\sigma_k^{(m)})$, the allocation strategy (σ, p) has value $V(\sigma,p) \leq 0$. This result is shown in Theorem B.7.

Step B.3 By combining Step B.1 and Step B.2, for any p, with the choice of time allocation sequences $\sigma = (\sigma_k^{(m)})$ considered above, we have

$$V(\sigma,p) = 0. \tag{4.12}$$

We consider $C^{\rho}(n) = \beta^n\Gamma^{(\rho_{n-1})}(t_n^{\rho})$ as a (sub-)compensator in Definition 3.21. The strategy ρ^* given in Theorem 4.3 is the strategy of always playing the bandit with the minimal index. Therefore, it lies in the same equivalence class as a strategy with the time allocation sequences $(\sigma^{(m)})$ (and with p indicating the minimum index amongst all bandits at each time). Hence, by (4.12),

$$\mathfrak{E}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho^*_{n-1})}(t_n^{\rho^*}) - \Gamma^{(\rho^*_{n-1})}(t_n^{\rho^*})\right)\right) = 0.$$

Furthermore, by (4.10),

$$\mathfrak{E}\left(\sum_{n=1}^{\infty}\beta^n\left(h^{(\rho_{n-1})}(t_n^{\rho}) - \Gamma^{(\rho_{n-1})}(t_n^{\rho})\right)\right) \geq 0.$$

By monotonicity of the process Γ, we prefer lower values earlier, due to the discount effect. Thus, we prove the optimality condition when $N = 0$. We can now restart our analysis at the considered (orthant) time to obtain the optimal condition for $N > 0$. We now thus show that ρ^* satisfies the condition for C-optimal. The formal proof of this result can be found in Theorem B.8.

5 Numerical results

In this section, we study the behaviour of the robust Gittins index using a numerical example. Again we omit the superscript (m) for notational simplicity.

We suppose the bandit under consideration generates independent identically distributed costs $(h(t))_{1\leq t\leq T}$ of either \$1 or \$0, given (unknown) probability $\mathbb{P}(h(t)=1)=\theta$ and $h(t)=2$ for all $t>T$. The filtration $(\mathcal{F}_t)_{t\geq 0}$ is generated by the observed cost process $(\xi_t)_{1\leq t\leq T}=(h(t))_{1\leq t\leq T}$ (with $\mathcal{F}_0$ trivial). The horizon T can be thought of as the maximum number of times that each bandit can be played.

Remark 5.1. An imaginary horizon T is introduced in order to allow us to easily construct a data-driven recursive nonlinear expectation (5.1) by backward induction

The future cost $h(t)=2$ is introduced to simplify our numerical method. By considering (4.1), we can see that the robust Gittins index $(\gamma(t))_{t\geq 1}$ takes values between 0 and 1 when $t\leq T$ and $\gamma(t)=2$ for $t>T$. Moreover, the optimal stopping time $\sigma(t,\gamma(t))\leq T-t$. 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 $\mathcal{E}_{(t)}(\cdot):L^\infty(\mathcal{F}_{t+1})\to L^\infty(\mathcal{F}_t)$. Once we have a one-step coherent nonlinear expectation, we can construct an $(\mathcal{F}_t)$-consistent coherent nonlinear expectation by

$$\mathcal{E}\Big(\ \cdot\ \Big|\mathcal{F}_t\Big)=\mathcal{E}_{(t)}\Big(\mathcal{E}_{(t+1)}\big(\cdots\mathcal{E}_{(T-1)}(\ \cdot\ )\cdots\big)\Big).$$

Remark 5.2. We will consider one-step coherent nonlinear expectation which is inspired by the DR-Expectation [23], see also Bielecki, Chen and Cialenco [14]):

$$\mathcal{E}_{(t)}\Big(f(\xi_1,...,\xi_t,\xi_{t+1})\Big):=\sup_{\theta\in\Theta_t}\Big(\theta f(\xi_1,...,\xi_t,1)+(1-\theta)f(\xi_1,...,\xi_t,0)\Big) \tag{5.1}$$

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 (possibly 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 choose a credible level $k\in[0,1]$ and obtain $p^\pm(p_t,n_t)$ by

$$p^\pm(p_t,n_t)=I^{-1}_{(p_tn_t,(1-p_t)n_t)}(0.5\pm k/2)$$

where $q\mapsto I^{-1}_{(a,b)}(q)$ is the quantile function of the Beta(a,b) distribution.

One could also use the central limit theorem to obtain an asymptotic confidence interval. However, due to the fact that $\Theta_t\subseteq[0,1]$, we restrict ourselves to the credible set above to avoid end-effects, and allow for asymmetry in the plausible values around the 'best' estimate.

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 $g_{k,T-t}:\mathbb{R}^2\to\mathbb{R}$ such that $\mathcal{E}\big(f(\xi_{t+1},...,\xi_T)\big|\mathcal{F}_t\big)=g_{k,T-t}\Big(p_t,\frac{1}{\sqrt{n_t}}\Big)$ where we write the nonlinear expectation as a function of $n_t^{-1/2}$ instead of n_t to approximate our function on a compact domain. The choice of $n_t^{-1/2}$ comes from the natural scaling of the credible set.

By recalling the definition of $\gamma(s)$ (Definition 4.1), one can show (using a general robust dynamic programming argument, as in Ruszczyński [71], or the nonlinear Snell's envelope, as in Riedel [68]) that we can write

$$\gamma(t)=\gamma_{k,\beta,T-t}\Big(p_t,\frac{1}{\sqrt{n_t}}\Big)\qquad\text{where}\qquad n_t=n_0+t.$$

for some function $\gamma_{k,\beta,T-t}$.

We then use a simple finite-difference algorithm (see Appendix D) to estimate the function

$$\Big(p, \frac{1}{\sqrt{n}}\Big) \mapsto \gamma_{k,\beta,T-(n-n_0)}\Big(p, \frac{1}{\sqrt{n}}\Big) - p$$

where, in our simulations, we fix $n_0 = 1$.

Plots of this estimate, for various values of k, β and T, can be found in Figure 1.

For $h(t) = \xi_t$, with uncertainty modeled by (5.1), at each time step we wish to play the bandit with the lowest θ. Classically, this is estimated by p, so a naïve (greedy) strategy would suggest playing the bandit with the lowest estimated average loss p. By using C-optimality, at each point, we choose a bandit 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 bandits.

In Figure 1, we see the following broad phenomena:

- 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 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 k is increased, the difference $\gamma - p$ increases. This is due to the fact that k corresponds to the 'width' of the 'credible interval'. Hence, large k means that we become more conservative and favour exploiting over exploring.

5.1 Prospect theory

One result suggested in Figure 1 when $\beta = 0.9999$ and $k = 0.01$ 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 $k = 0.95$, or $\beta = 0.95$ and $k = 0.8$, we become more pessimistic.

Curiously, when $\beta = 0.9999$ and $k = 0.8$, or $\beta = 0.95$ and $k = 0.5$, 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 [47] or in rank-dependent expected utility by Quiggin [65, 66]. 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 k) 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.

5.2 Monte-Carlo simulation

In order to illustrate the performance of the robust Gittins index calculated above in real decision making, we consider the Bernoulli bandit as described above over 50 exchangeable bandits and for a horizon $T = 10^4$. We run 10^3 Monte-Carlo simulations and compare 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

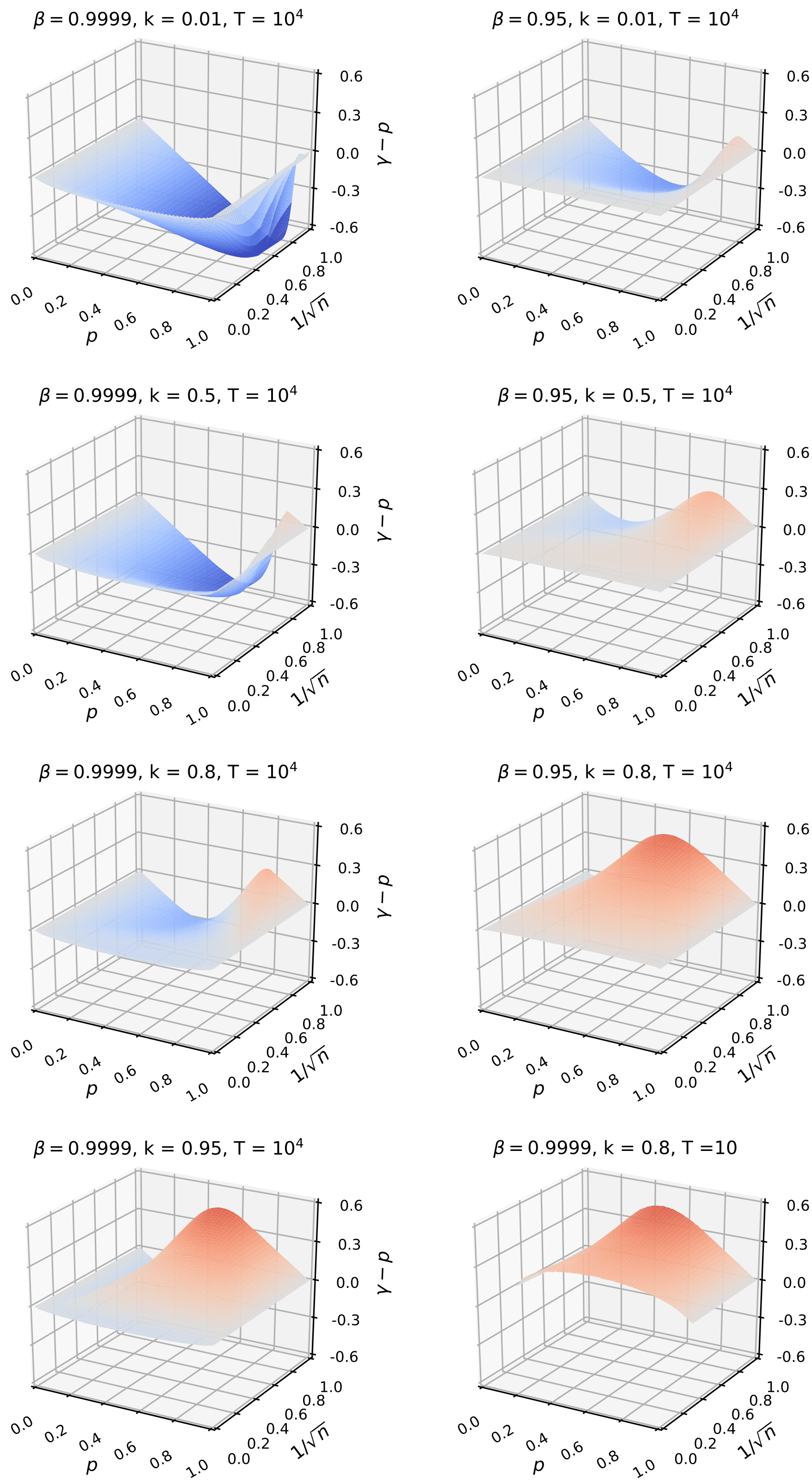

Figure 1: Estimated value of $\gamma - p$ for different values of k, β and T. The case $T = 10$ is truncated as n cannot exceed T (by definition).

a, b independently from a $\Gamma(1, 1/100)$ distribution, then generate the 'true' probabilities for each bandit independently from Beta(a, b). We generate 10 trials on each bandit to provide initial information.

N.B. Formally, we assume that each bandit can be played for at least $T = 10^4$ trials in constructing our Gittins index. We illustrate the performance of the first 10^4 plays to compare with other algorithms.

5.2.1 Measures of regret

There are a number of possible objectives to measure the loss of our decisions. We will consider the following examples (from Bertini et al. [19] and Lai and Robbins [54]).

- **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 $R(L) = \sum_{n=0}^{L} (\theta^{(\rho_n)} - \theta^*)$ where $\theta^{(m)}$ is the true probability of the mth bandit and $\theta^* = \min_m \theta^{(m)}$.
- **Sub-optimal plays.** This measures the number of times where we play a sub-optimal bandit which is given by $N_\vee(L) = \sum_{n=0}^{L} \mathbb{I}(\theta^{(\rho_n)} \neq \theta^*)$.

5.2.2 Policy for multi-armed-bandits

In our simulation, we will label our algorithm the DR (Data-Robust) algorithm. We also consider the following classical policies which are commonly used to solve the Bernoulli bandit problem. These policies choose an arm by considering the minimal index I evaluated on each bandit separately. Literature about these policies and further developments can be found in the reviews by Bertini et al. [19] or Russo et al. [70]. For notational simplicity, we will denote by p and n the estimated probability and the number of observations of the considered bandit at the time before making a decision.

- **Greedy strategy.** In this policy, we choose the bandit with the minimal estimated probability given by $I^{Greedy} = p$.
- **Thompson strategy.** This is a Bayesian adaptive decision strategy for the bandit problem. It proceeds by first randomly generating a sample from the posterior distribution of the mean cost of each bandit, then chooses to play the bandit which gave the minimal sample. In our setting, these samples are given by $I^{Thompson} \sim$ Beta$(a_0 + pn, b_0 + (1-p)n)$ where $a_0, b_0 > 0$ are the parameters of a Beta prior distribution for the mean. To avoid biasing our estimation, we consider initial values $a_0 = b_0 \in \{0.0001, 1, 50\}$, where larger values correspond to a more informative prior.
- **UCB strategy.** This is an optimistic strategy to choose the bandit based on its lower bound. $I^{UCB} = p - \sqrt{\frac{\lambda \log N}{n}}$ where $\lambda > 0$ is a chosen parameter, which is commonly chosen to be 2 and $N := \sum_{m=1}^{M} n^{(m)}$ is the total number of observations across all bandits.

Remark 5.3. To avoid bias in the algorithms, we choose a bandit uniformly at random if there is more than one bandit with minimal index.

In Figure 2, considering first the cases where $\beta = 0.9999$, we can see that an increase in the value of k has a nonlinear effect on the distribution of regret. Initially, increasing k appears to lead to a reduction in the typical regret, but a possible increase in the average and variability of the number of suboptimal plays. However, setting k too large clearly leads to worse outcomes. This is because k 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

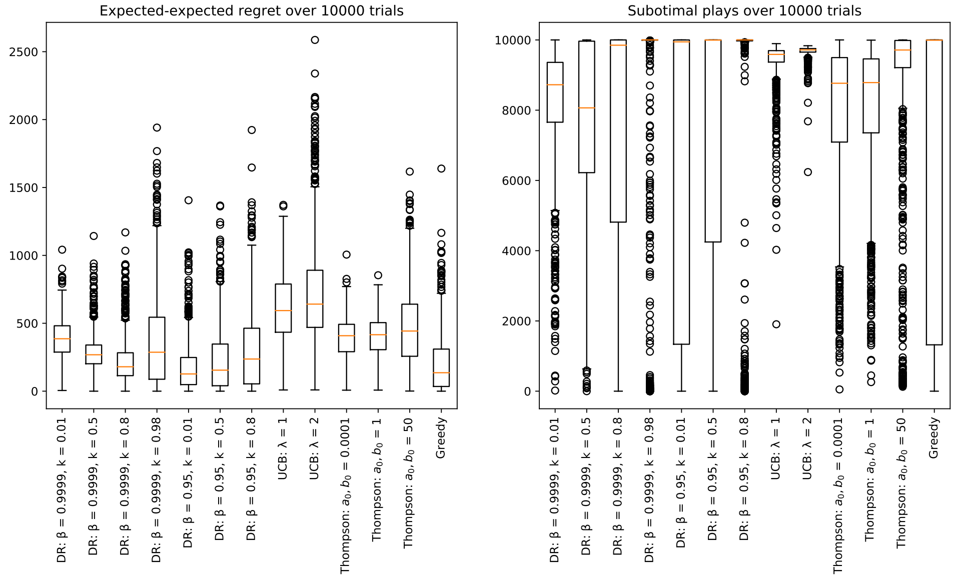

Figure 2: Regret under different policies.

exploit. It follows that a large value of k encourages us to exploit early, and we may not find the optimal bandit 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 k) 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 $k = 0.8$. In this case the average expected-expected regret is relatively small when compared to other strategies, but its average number of suboptimal plays is relatively high. Reducing β to 0.95 emphasizes these effects even further.

As discussed in the introduction, the UCB algorithm asymptotically achieves a minimal regret bound (see [54]). It does so by ensuring that, over short horizons, the algorithm explores a sufficient amount, in order to guarantee good asymptotic performance. We can see that in our simulation (with 10^4 plays over 50 bandits), the UCB algorithm is still in its high exploration regime which results in a high regret and very few optimal plays. In contrast, a Greedy algorithm always chooses an arm to play without taking into account its uncertainty (and so without considering the possibility for exploration) and therefore there is no learning in its procedure. This results in the greedy algorithm yielding a low average regret but a high average number of suboptimal plays.

5.2.3 Robustness of the DR algorithms

In Figure 3, we illustrate the interquartile range and the standard deviation of the total expected-expected regret when $\beta = 0.9999$ with different values of k over 1000 simulations. We can see that by introducing an appropriate value of k, we can obtain a substantial reduction in the interquartile range and the standard deviation. In particular, the DR algorithm not only gives a low average regret, but does so consistently over different simulations.

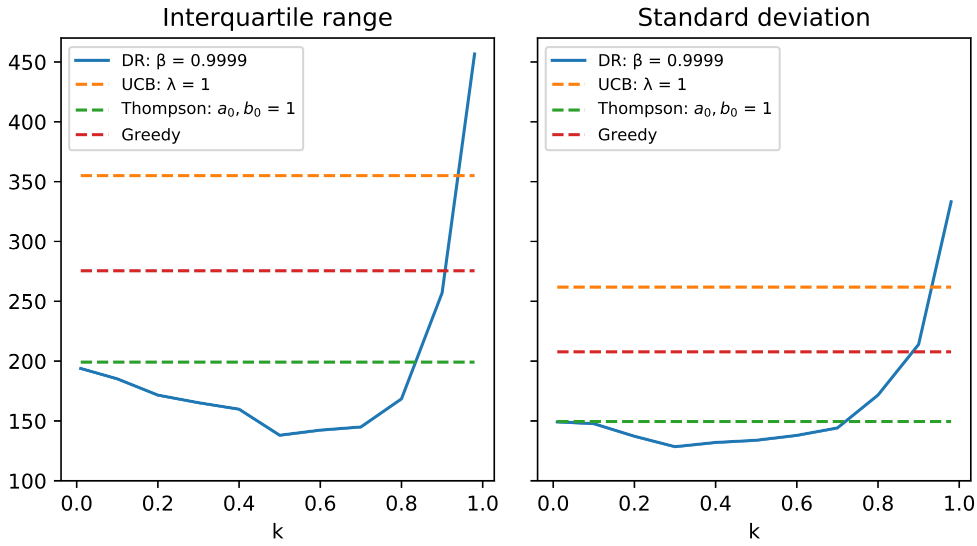

Figure 3: Deviation of the expected-expected regret when $\beta = 0.9999$.

A Part A: analysis of a single bandit

We will now flesh out the sketch given in Section 4.2. In this section, we focus our discussion on a single bandit.

A.1 Step A.1: indifference reward and optimal stopping problem

We first recall the definition of the robust Gittins index (process).

$$\gamma(s) := \operatorname{ess\,inf} \left\{ \gamma \in L^\infty(\mathcal{F}_s) : \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} \mathcal{E}\left(\sum_{t=s+1}^{s+\tau} \beta^t \left(h(t) - \gamma\right) \,\middle|\, \mathcal{F}_s \right) \leq 0 \right\}$$

where $\mathcal{T}(s)$ denotes the family of $(\mathcal{F}_{s+t})_{t\geq 0}$-positive stopping times.

Remark A.1. If $\tau = 1$, it follows from boundedness of h (Assumption 3.1) that $\gamma(t) < C$.

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 A.2. *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 A.3. *The function V_s defined above satisfies* $\operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \gamma(s)) = 0$.

Proof. This can be done by showing that V_s satisfies the regularity assumptions of Lemma C.6.

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

$$\begin{aligned}\Big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \lambda\big)\Big) &= \Big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \lambda'\big)\Big) + \sum_{t=s+1}^{s+\tau} \beta^t (\lambda' - \lambda) \\ &\leq \Big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \lambda'\big)\Big) + \Big(\frac{\beta^{s+1}}{1-\beta}\Big)(\lambda' - \lambda).\end{aligned}$$

By monotonicity and translation equivariance, we have

$$\begin{aligned}0 \leq V_s(\tau, \lambda) - V_s(\tau, \lambda') &= \mathcal{E}\Big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \lambda\big) \,\Big|\, \mathcal{F}_s\Big) - \mathcal{E}\Big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \lambda'\big) \,\Big|\, \mathcal{F}_s\Big) \\ &\leq \Big(\frac{\beta^{s+1}}{1-\beta}\Big)(\lambda' - \lambda).\end{aligned}$$

So, V_s is Lipschitz in λ.

Condition (iii) follows from $(\mathcal{F}_t)_{t\geq 0}$-regularity of $\mathcal{E}$ (Remark 3.4). The result follows from Lemma C.6. □

Corollary A.4. *For every $\tau \in \mathcal{T}(s)$, we have $\mathcal{E}\big(\sum_{t=s+1}^{s+\tau} \beta^t \big(h(t) - \gamma(s)\big) \big| \mathcal{F}_s\big) \geq 0$.*

Remark A.5. Theorem A.3 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 bandit.

A.2 Step A.2: optimal stopping time

By considering a Snell envelope argument, as in Riedel [68] with slight modification, we can establish that 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 C.7). In this subsection, we will show that τ^* can be expressed as a hitting time of the Gittins index process $(\gamma(s))$.

Definition A.6. *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 A.5, 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 bandit attractive so, to minimize the total 'expected' cost, we will stop.

In what follows, we formalize 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 A.3).

The optimality of $\sigma(s, \lambda)$ can be proved by showing that for any stopping time $\tau \in \mathcal{T}(s)$, if $\tau > \sigma(s, \lambda)$, our value can be improved by stopping at $\sigma(s, \lambda)$ (Lemma A.7). On the other hand, if $\tau < \sigma(s, \lambda)$, the value can be improved by continuing to play (Lemma A.8).

Lemma A.7. *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 A.4 together with time-consistency and monotonicity of our nonlinear expectation.

Define $\nu = \tau \wedge \sigma(s,\lambda)$. By Corollary A.4 and regularity (Remark 3.4),

$$\begin{aligned}
0 &\le \mathcal{E}\Big(\sum_{t=s+\nu+1}^{s+\tau} \beta^t\left(h(t)-\gamma(s+\nu)\right)\Big|\mathcal{F}_{s+\nu}\Big)\\
&= \mathbb{I}(\tau>\sigma(s,\lambda))\mathcal{E}\Big(\sum_{t=s+\nu+1}^{s+\tau} \beta^t\left(h(t)-\gamma(s+\sigma(s,\lambda))\right)\Big|\mathcal{F}_{s+\nu}\Big)+\mathbb{I}(\tau\le\sigma(s,\lambda))(0)\\
&\le \mathbb{I}(\tau>\sigma(s,\lambda))\mathcal{E}\Big(\sum_{t=s+\nu+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\nu}\Big)=\mathcal{E}\Big(\sum_{t=s+\nu+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\nu}\Big).
\end{aligned}$$

By translation equivariance,

$$\begin{aligned}
\sum_{t=s+1}^{s+\nu} \beta^t\left(h(t)-\lambda\right) &\le \sum_{t=s+1}^{s+\nu} \beta^t\left(h(t)-\lambda\right)+\mathcal{E}\Big(\sum_{t=s+\nu+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\nu}\Big)\\
&= \mathcal{E}\Big(\sum_{t=s+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\nu}\Big).
\end{aligned}$$

By monotonicity and time-consistency,

$$\mathcal{E}\Big(\sum_{t=s+1}^{s+\nu} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s}\Big)\le \mathcal{E}\Big(\mathcal{E}\Big(\sum_{t=s+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\nu}\Big)\Big|\mathcal{F}_s\Big)=\mathcal{E}\Big(\sum_{t=s+1}^{s+\tau} \beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s}\Big).$$

In particular, $V_s(\tau\wedge\sigma(s,\lambda),\lambda)=V_s(\nu,\lambda)\le V_s(\tau,\lambda)$. □

Lemma A.8. *Let $\tau\in\mathcal{T}(s)$ and let $\lambda\in L^\infty(\mathcal{F}_s)$ taking values in $[0,C)$. Then there exists a stopping time $\tau_1\in\mathcal{T}(s)$ with $\tau_1\ge\tau$ such that $V_s(\tau,\lambda)\ge V_s(\tau_1,\lambda)$ and on the event $A:=\{\gamma(s+\tau)\le\lambda\}$, we have $\tau_1>\tau$.*

Proof. For $A=\{\gamma(s+\tau)\le\lambda\}$, define $\gamma^\tau:=\gamma(s+\tau)\mathbb{I}_{A^c}+\lambda\mathbb{I}_A\ge\gamma(s+\tau)$. By Theorem A.3 and monotonicity of our nonlinear expectation,

$$0=\operatorname*{ess\,inf}_{\tilde\tau\in\mathcal{T}(s+\tau)}\mathcal{E}\Big(\sum_{t=s+\tau+1}^{s+\tau+\tilde\tau}\beta^t\left(h(t)-\gamma(s+\tau)\right)\Big|\mathcal{F}_{s+\tau}\Big)\ge\operatorname*{ess\,inf}_{\tilde\tau\in\mathcal{T}(s+\tau)}\mathcal{E}\Big(\sum_{t=s+\tau+1}^{s+\tau+\tilde\tau}\beta^t\left(h(t)-\gamma^\tau\right)\Big|\mathcal{F}_{s+\tau}\Big).$$

Thus, by Theorem C.7, there exists $\tilde\tau^*\in\mathcal{T}(s+\tau)$ such that,

$$0\ge\mathcal{E}\Big(\sum_{t=s+\tau+1}^{s+\tau+\tilde\tau^*}\beta^t\left(h(t)-\gamma^\tau\right)\Big|\mathcal{F}_{s+\tau}\Big).$$

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 3.4), it follows that

$$\begin{aligned}
\mathcal{E}\Big(\sum_{t=s+1}^{s+\tau_1}\beta^t\left(h(t)-\lambda\right)\Big|\mathcal{F}_{s+\tau}\Big) &= \sum_{t=s+1}^{s+\tau}\beta^t\left(h(t)-\lambda\right)+\mathbb{I}_A\mathcal{E}\Big(\sum_{t=s+1}^{s+\tau+\tilde\tau^*}\beta^t\left(h(t)-\gamma^\tau\right)\Big|\mathcal{F}_{s+\tau}\Big)\\
&\ge \sum_{t=s+1}^{s+\tau}\beta^t\left(h(t)-\lambda\right).
\end{aligned}$$

Finally, by applying monotonicity and time-consistency as in the previous lemma, the result follows. □

Corollary A.9. *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*

$$V_s(\tau, \lambda) \geq V_s(\tau_1, \lambda) \geq ... \geq V_s(\tau_n, \lambda)$$

and on the event $\bigcap_{k=1}^{n-1} \{\gamma(s+\tau_k) \leq \lambda\}$ we have $\tau_n > \tau_{n-1} > ... > \tau_1 > \tau$. 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 A.10. *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 A.4.

Proof. By Lemma A.7 and Corollary A.9,

$$V_s(\tau, \lambda) \geq V_s(\tau_n, \lambda) \geq V_s(\tau_n \wedge \sigma(s, \lambda), \lambda).$$

Observe that by Corollary A.9,

$$\{\tau_n < \sigma(s, \lambda)\} = \{\gamma(s+\theta) \leq \lambda \ \ \forall \theta \leq \tau_n\} \subseteq \bigcap_{k=1}^{n-1} \{\gamma(s+\tau_k) \leq \lambda\} \subseteq \{\tau_n \geq n\}.$$

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

$$\mathcal{E}\bigg(\sum_{t=s+1}^{s+\tau_n \wedge \sigma(s,\lambda)} \beta^t \big(h(t) - \lambda\big) \ \bigg| \ \mathcal{F}_s \bigg) \longrightarrow \mathcal{E}\bigg(\sum_{t=s+1}^{s+\sigma(s,\lambda)} \beta^t \big(h(t) - \lambda\big) \ \bigg| \ \mathcal{F}_s \bigg).$$

In particular, $V_s(\tau, \lambda) \geq V_s(\sigma(s, \lambda), \lambda)$. □

Remark A.11. Bank and El Karoui [9] consider a similar result to this theorem, 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 [10] and [11] for further discussion).

A.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 bandit. 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 analyze our (minimal) compensation process.

Definition A.12. *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 $\hat{S}_0 := 0$.

(ii) Given $\hat{S}_n$, define $\sigma_n := \inf\{\theta \geq 1 \, : \, \gamma(\hat{S}_n + \theta) > \gamma(\hat{S}_n)\}$ and $\hat{S}_{n+1} := \hat{S}_n + \sigma_n$.

Equivalently, we can define $\sigma_n := \sigma(\hat{S}_n, \gamma(\hat{S}_n))$ as in Definition A.6.

Definition A.13. *We define the* prevailing reward *process Γ by the running maximum of γ, that is, $\Gamma(t) := \max_{0\leq\theta\leq t-1}\gamma(\theta)$.*

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 A.14. *For all $n \in \mathbb{N}$, $\mathcal{E}\big(\sum_{t=\hat{S}_n+1}^{\infty}\beta^t\,(h(t)-\Gamma(t))\,\big|\mathcal{F}_{\hat{S}_n}\big) = 0$. In particular,*

$$\mathcal{E}\bigg(\sum_{t=1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\bigg) = 0.$$

Proof. By Theorem A.10, for all $k \in \mathbb{N}$, $0 = \mathcal{E}\big(\sum_{t=\hat{S}_k+1}^{\hat{S}_{k+1}}\beta^t\big(h(t)-\gamma(\hat{S}_n)\big)\big|\mathcal{F}_{\hat{S}_k}\big)$. Now, fix $n, N \in \mathbb{N}$ with $N \geq n$. By time-consistency and translation equivariance,

$$\begin{aligned}
&\mathcal{E}\bigg(\sum_{t=\hat{S}_n+1}^{\hat{S}_N}\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_{\hat{S}_n}\bigg)\\
&= \mathcal{E}\bigg(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}}\beta^t\big(h(t)-\Gamma(t)\big) + \mathcal{E}\bigg(\sum_{t=\hat{S}_{N-1}+1}^{\hat{S}_N}\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_{\hat{S}_{N-1}}\bigg)\bigg|\mathcal{F}_{\hat{S}_n}\bigg)\\
&= \mathcal{E}\bigg(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}}\beta^t\big(h(t)-\Gamma(t)\big) + \mathcal{E}\bigg(\sum_{t=\hat{S}_{N-1}+1}^{\hat{S}_N}\beta^t\big(h(t)-\gamma(\hat{S}_{N-1})\big)\bigg|\mathcal{F}_{\hat{S}_{N-1}}\bigg)\bigg|\mathcal{F}_{\hat{S}_n}\bigg)\\
&= \mathcal{E}\bigg(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-1}}\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_{\hat{S}_n}\bigg) = \mathcal{E}\bigg(\sum_{t=\hat{S}_n+1}^{\hat{S}_{N-2}}\beta^t\big(h(t)-\Gamma(t)\big)\bigg|\mathcal{F}_{\hat{S}_n}\bigg) = \cdots = 0.
\end{aligned}$$

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 the Lebesgue property, the result follows. □

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 A.15. *Let $\tau \in \mathcal{T}(s)$ with $1 \leq \tau \leq \sigma := \sigma(s,\lambda)$. Then*

$$\mathcal{E}\bigg(\sum_{t=s+\tau+1}^{s+\sigma}\beta^t\,(h(t)-\lambda)\,\bigg|\mathcal{F}_{s+\tau}\bigg) \leq 0.$$

Proof. Write $H_\tau := \mathcal{E}\left(\sum_{t=s+\tau+1}^{s+\sigma}\beta^t\Big(h(t)-\lambda\Big)\Big|\mathcal{F}_{s+\tau}\right)$ and $A := \{H_\tau > 0\}$.

Define $\tilde{\sigma} := \tau\,\mathbb{I}_A + \sigma\,\mathbb{I}_{A^c}$.

$$\begin{aligned}\mathcal{E}\Big(\sum_{t=s+1}^{s+\sigma}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_{s+\tau}\Big) &= \sum_{t=s+1}^{s+\tau}\beta^t\,(h(t)-\lambda) + \mathcal{E}\Big(\sum_{t=s+\tau+1}^{s+\sigma}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_{s+\tau}\Big)\\ \geq \sum_{t=s+1}^{s+\tau}\beta^t\,(h(t)-\lambda) + \mathcal{E}\Big(\sum_{t=s+\tau+1}^{s+\sigma}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_{s+\tau}\Big)\mathbb{I}_{A^c} &= \mathcal{E}\Big(\sum_{t=s+1}^{s+\tilde{\sigma}}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_{s+\tau}\Big).\end{aligned}$$

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

$$\mathcal{E}\Big(\sum_{t=s+1}^{s+\sigma}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_s\Big) > \mathcal{E}\Big(\sum_{t=s+1}^{s+\tilde{\sigma}}\beta^t\,(h(t)-\lambda)\Big|\mathcal{F}_s\Big).$$

This contradicts the minimality of $\sigma(s,\lambda)$ established in Theorem A.10. □

Theorem A.16. *For all $N \in \mathbb{N}$,*

$$\mathcal{E}\Big(\sum_{t=N+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_N\Big) \leq 0.$$

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 A.14 and Lemma A.15,

$$\begin{aligned}&\mathcal{E}\Big(\sum_{t=\tau_n+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_{\tau_n}\Big)\\ &\quad= \mathcal{E}\Big(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}}\beta^t\big(h(t)-\Gamma(t)\big) + \mathcal{E}\Big(\sum_{t=\hat{S}_{n+1}+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_{\hat{S}_{n+1}}\Big)\Big|\mathcal{F}_{\tau_n}\Big)\\ &\quad= \mathcal{E}\Big(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_{\tau_n}\Big) \;=\; \mathcal{E}\Big(\sum_{t=\tau_n+1}^{\hat{S}_{n+1}}\beta^t\big(h(t)-\gamma(\hat{S}_n)\big)\Big|\mathcal{F}_{\tau_n}\Big) \leq 0.\end{aligned}$$

Therefore, as $\{\hat{S}_n \leq N < \hat{S}_{n+1}\}$ is $\mathcal{F}_N$-measurable, by the Lebesgue property and regularity (Remark 3.4),

$$\begin{aligned}&\mathcal{E}\Big(\sum_{t=N+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_N\Big)\\ &\quad= \mathcal{E}\Big(\lim_{L\to\infty}\Big(\sum_{n=0}^{L}\mathbb{I}_{\{\hat{S}_n\leq N<\hat{S}_{n+1}\}}\Big)\Big(\sum_{t=\tau_n+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big)\Big|\mathcal{F}_N\Big)\\ &\quad= \lim_{L\to\infty}\sum_{n=0}^{L}\mathbb{I}_{\{\hat{S}_n\leq N<\hat{S}_{n+1}\}}\mathcal{E}\Big(\sum_{t=\tau_n+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_N\Big)\\ &\quad= \lim_{L\to\infty}\sum_{n=0}^{L}\mathbb{I}_{\{\hat{S}_n\leq N<\hat{S}_{n+1}\}}\mathbb{I}_{\{\tau_n\geq N\}}\mathcal{E}\Big(\sum_{t=\tau_n+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_N\Big)\\ &\quad= \lim_{L\to\infty}\sum_{n=0}^{L}\mathbb{I}_{\{\hat{S}_n\leq N<\hat{S}_{n+1}\}}\mathbb{I}_{\{\tau_n\geq N\}}\mathcal{E}\Big(\mathcal{E}\Big(\sum_{t=\tau_n+1}^{\infty}\beta^t\big(h(t)-\Gamma(t)\big)\Big|\mathcal{F}_{\tau_n}\Big)\Big|\mathcal{F}_N\Big) \leq 0. \quad\square\end{aligned}$$

Remark A.17. 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.

A.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 A.16) until the horizon (i.e. the total expected loss is zero, as in Proposition A.14). 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 bandits.

Theorem A.18. *By Assumption 3.10, recall that $\mathcal{E}$ admits a robust representation of the form $\mathcal{E}\big(\,\cdot\,\big) = \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}(\cdot)$.*

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

$$\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=1}^{\infty} \alpha(t)\beta^t\big(h(t) - \Gamma(t)\big)\bigg) \geq -\epsilon.$$

Proof. By Proposition A.14 and the robust representation theorem, for a fixed $\epsilon > 0$, we can find a probability measure $\mathbb{Q} \in \mathcal{Q}$ such that $\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\big) \geq -\epsilon$.

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

$$\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=1}^{\infty} \alpha^N(t)\beta^t\big(h(t) - \Gamma(t)\big)\bigg) \geq -\epsilon. \tag{A.1}$$

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

$$-\epsilon \leq \mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=1}^{N} \alpha(t)\beta^t\big(h(t) - \Gamma(t)\big)\bigg) + \mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{T} \alpha(N)\beta^t\big(h(t) - \Gamma(t)\big)\bigg). \tag{A.2}$$

By Theorem 3.7, $\mathcal{E}\big(\,\cdot\,\big|\mathcal{F}_N\big) = \operatorname{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\big(\,\cdot\,\big|\mathcal{F}_N\big)$. Combining this with Theorem A.16, we can see that

$$\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\bigg|\mathcal{F}_N\bigg) \leq 0.$$

Since α is decreasing,

$$\big(\alpha(N) - \alpha(N+1)\big)\mathbb{E}^{\mathbb{Q}}\bigg(\sum_{t=N+1}^{\infty} \beta^t\big(h(t) - \Gamma(t)\big)\bigg|\mathcal{F}_N\bigg) \leq 0.$$

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 (A.2), we prove (A.1) with N replaced by $N+1$ and done the induction step. Finally, by the bounded convergence theorem, we can take $N \to \infty$ on (A.1) and obtain the required result. □

B Part B: analysis of multiple bandits

We are now ready to consider the problem of choosing between multiple bandits.

In Definition 4.10, we introduce our class of admissible controls which can be considered in our dynamic allocation problem. This class of controls introduces a few natural ways of parameterizing time. We therefore will use the following terminology to describe the evolution of time 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 bandits. (This corresponds to the time parameter of the simple form ρ which is briefly discussed earlier on in Remark 2.7 and later in Definition 4.12.)

2. **'Trial'** refers to the number of times that we play a specific bandit. (This corresponds to the sum $\sum_{k=0}^{K} \tau_k^{(m)}$.)

3. **'Decision'** refers to the number of times that we make a decision between bandits. (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 bandit. (This corresponds to the time parameter of the allocation sequence $(\tau_k^{(m)})_{k\geq 0}$ for each fixed $m \in \mathcal{M}$.)

Remark B.1. 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 4.7).

Remark B.2. The mth component of the recording sequence η_n (Definition 4.8) represents the number of runs in the mth bandit before the nth decision. We can see that the random variable η_n takes values in $\left\{r \in \mathcal{S} : \sum_{m=1}^{M} r^{(m)} \leq n\right\}$.

In order to prove C-optimality, we consider the target function as stated in (4.9).

Definition B.3. *For each $m \in \mathcal{M}$, let $\left(h^{(m)}(t)\right)_{t\geq 1}$ be the uniformly bounded non-negative cost process at the tth trial of the mth bandit with prevailing reward process $\left(\Gamma^{(m)}(t)\right)_{t\geq 1}$ (Definition A.13). For an allocation strategy (τ, p), (Definition 4.10), we define the* Gittins target function *by*

$$V(\tau, p) := \mathfrak{E}\left(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\right)\right)$$

where ρ is the simple form of (τ, p) with corresponding counting processes $t_n^\rho := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = \rho_{n-1})$, and $\mathfrak{E}$ is a partially consistent orthant nonlinear expectation, as in Definition 3.11.

Remark B.4. 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 B.7 in Step B.2. This definition, however, makes it clear that V depends on (τ, p) only through its simple form.

B.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 B.5. *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 $\mathcal{H}' := \{\emptyset, H\}$. 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 B.6. *For any allocation strategy (τ, p), we have $V(\tau, p) \geq 0$.*

Proof. Let ρ be the simple form of (τ, p). Write $R_t^{(m)}$ for the total number of trials on other bandits before making the tth trial on the mth bandit, i.e.

$$R_t^{(m)} := \sum_{k\neq m} \sum_{n=0}^{N_t^{(m)}} \mathbb{I}(\rho_n = k) \quad \text{where} \quad N_t^{(m)} := \inf\Big\{N \geq 0 : \sum_{n=0}^{N} \mathbb{I}(\rho_n = m) = t\Big\}.$$

Since $R_t^{(m)}$ does not depend on future realizations of the mth bandit, $R_t^{(m)}$ is $\mathcal{F}_{t-1}^{(m)} \otimes \big(\bigotimes_{k=1,k\neq m}^{M} \mathcal{F}_\infty^{(k)}\big)$-measurable (taking the product in an appropriate order). Moreover, as $N_t^{(m)}$ is increasing in t, it follows that $R_t^{(m)}$ is increasing in t.

Now, fix $\epsilon > 0$. By Theorem A.18, for each $m \in \mathcal{M}$, we can find a probability measure $\mathbb{Q}^{(m)} \in \mathcal{Q}^{(m)}$ such that, for every adapted decreasing process $(\alpha^{(m)}(t))_{t\geq 0}$ taking values in $[0, 1]$, we have

$$\mathbb{E}^{\mathbb{Q}^{(m)}}\Big(\sum_{t=1}^{\infty} \alpha^{(m)}(t)\beta^t\big(h^{(m)}(t) - \Gamma^{(m)}(t)\big)\Big) \geq -\epsilon. \tag{B.1}$$

Define $\tilde{\alpha}^{(m)}(t) := \int_{\prod_{k\neq m} \Omega^{(k)}} \beta^{R_t^{(m)}} \, d\big(\bigotimes_{k\neq m} \mathbb{Q}^{(k)}\big)$. By Fubini's theorem, as $R_t^{(m)}$ is $\mathcal{F}_{t-1}^{(m)} \otimes \big(\bigotimes_{k=1,k\neq m}^{M} \mathcal{F}_\infty^{(k)}\big)$-measurable and $\beta \in (0, 1]$, the process $\big(\tilde{\alpha}^{(m)}(t)\big)$ is an $\big(\mathcal{F}_t^{(m)}\big)$-predictable process taking values in $[0, 1]$.

Moreover, $R_t^{(m)}$ is $\mathcal{F}_\infty^{(m)} \otimes \big(\bigotimes_{k=1,k\neq m}^{M} \mathcal{F}_\infty^{(k)}\big)$-measurable, so by Corollary B.5,

$$\tilde{\alpha}^{(m)}(t) = \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(k)}}\Big(\beta^{R_t^{(m)}} \,\Big|\tilde{\mathcal{F}}_\infty^{(m)}\Big) \quad \text{where} \quad \tilde{\mathcal{F}}_\infty^{(m)} := \mathcal{F}_\infty^{(m)} \otimes \bigotimes_{k\neq m}^{M} \mathcal{F}_0^{(k)}.$$

As $t \mapsto R_t^{(m)}$ is increasing, it then follows that $\tilde{\alpha}^{(m)}(t)$ is decreasing in t. Hence, by Theorem A.18 we obtain (B.1) with $\alpha^{(m)}$ replaced with $\tilde{\alpha}^{(m)}$.

By the definition of $\mathfrak{E}$ and Fubini's theorem, it follows that

$$\begin{aligned}
\mathfrak{E}\Big(\sum_{n=1}^{\infty} \beta^n \Big(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\Big)\Big) &\geq \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(m)}}\Big(\sum_{n=1}^{\infty} \beta^n \Big(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\Big)\Big) \\
&= \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(m)}}\Big(\sum_{m=1}^{M}\sum_{t=1}^{\infty} \beta^{R_t^{(m)}}\beta^t \Big(h^{(m)}(t) - \Gamma^{(m)}(t)\Big)\Big) \\
&= \sum_{m=1}^{M} \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(m)}}\Big(\sum_{t=1}^{\infty} \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(m)}}\Big(\beta^{R_t^{(m)}}\beta^t \Big(h^{(m)}(t) - \Gamma^{(m)}(t)\Big) \Big|\tilde{\mathcal{F}}_\infty^{(m)}\Big)\Big) \\
&= \sum_{m=1}^{M} \mathbb{E}^{\bigotimes_{k=1}^{M} \mathbb{Q}^{(m)}}\Big(\sum_{t=1}^{\infty} \tilde{\alpha}^{(m)}(t)\beta^t \Big(h^{(m)}(t) - \Gamma^{(m)}(t)\Big)\Big) \\
&= \sum_{m=1}^{M} \mathbb{E}^{\mathbb{Q}^{(m)}}\Big(\sum_{t=1}^{\infty} \tilde{\alpha}^{(m)}(t)\beta^t \Big(h^{(m)}(t) - \Gamma^{(m)}(t)\Big)\Big) \geq -M\epsilon.
\end{aligned}$$

As ϵ is arbitrary, the result follows. ∎

B.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 B.7. *For each $m \in \mathcal{M}$, let $(\sigma_k^{(m)})_{k\geq 0}$ be the sequence of running maximum random times associated to the mth bandit, as defined in Definition A.12, i.e. we define $\sigma_k^{(m)}$ and $\hat{S}_k^{(m)}$ recursively by $S_k^{(m)} := \sum_{l=0}^{k-1} \sigma_l^{(m)}$ and*

$$\sigma_k^{(m)} := \inf\left\{\theta \geq 1 : \gamma^{(m)}(S_k^{(m)} + \theta) > \gamma^{(m)}(S_k^{(m)})\right\}.$$

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.
$$\tilde{\Theta}_n := \sum_{m=1}^{M} \sum_{i=0}^{\eta_n^{(m)}-1} \sigma_i^{(m)}.$$
- $\tilde{\sigma}_n$ denotes the duration we decide to play following the nth decision time, i.e.
$$\tilde{\sigma}_n := \sigma_k^{(m)} \qquad \text{on the event } \{p_n = m, \eta_n^{(m)} = k\}.$$
N.B. $p_n = m$ means that we decide to play the mth bandit at the nth decision. The event $\eta_n^{(m)} = k$ means that we have had k runs of the mth bandit before the nth decision. Thus, we choose to make $\sigma_k^{(m)}$ more trials on this bandit before making another decision.
- $\tilde{\Psi}_n^{(m)}$ denotes the total number of trials on the mth bandit before making the nth decision, i.e.
$$\tilde{\Psi}_n^{(m)} := \sum_{i=0}^{k-1} \sigma_i^{(m)} \qquad \text{on the event } \{\eta_n^{(m)} = k\}.$$

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 the Lebesgue property of $\mathfrak{E}$, we can show that V agrees with Definition B.3 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 (4.7):

$$\Psi_r^{(m)} := \sum_{i=0}^{r^{(m)}-1} \sigma_i^{(m)} \qquad \text{and} \qquad \Theta_r := \sum_{m=1}^{M} \Psi_r^{(m)} \tag{B.2}$$

for

$$r \in \underline{\mathcal{S}}_N := \left\{ r \in \mathcal{S} : \sum_{m=1}^{M} r^{(m)} \le N,\ r^{(m)} \le T^{(m)} \right\}.$$

Note that $\Psi_r^{(m)} = \tilde{\Psi}_N^{(m)}$ and $\Theta_r = \tilde{\Theta}_N$ on the event $\{\eta_N = r\}$. From the definition of the recording sequence (Definition 4.8), it follows that Θ_r and the event $A_N^{(r,m)} := \{\eta_N = r, p_N = m\}$ are both $\mathcal{F}(\Psi_r)$-measurable.

On an event $A_N^{(r,m)}$, $\tilde{\Psi}_N^{(m)} = \Psi_r^{(m)}$ and $\Psi_r^{(m)}$ is a stopping time with respect to the filtration $(\mathcal{F}_t^{(m)})_{t\ge 0}$. By considering the optimality obtained in Theorem A.10, we can show that

$$\begin{aligned}
&\mathfrak{E}\bigg(\beta^{\Theta_N}\bigg(\sum_{l=1}^{\tilde{\sigma}_N}\beta^l\Big(h^{(p_N)}(\Psi_N^{(m)}+l)-\Gamma^{(p_N)}(\Psi_N^{(m)}+l)\Big)\bigg)\bigg)\\
&=\mathfrak{E}\bigg(\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathbb{I}_{A_N^{(r,m)}}\beta^{\Theta_r}\bigg(\sum_{l=1}^{\sigma_r(m)}\beta^l\Big(h^{(m)}(\Psi_r^{(m)}+l)-\Gamma^{(m)}(\Psi_r^{(m)}+l)\Big)\bigg)\bigg)\\
&=\mathfrak{E}\bigg(\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathbb{I}_{A_N^{(r,m)}}\beta^{\Theta_r}\bigg(\sum_{l=1}^{\sigma_r(m)}\beta^l\Big(h^{(m)}(\Psi_r^{(m)}+l)-\gamma^{(m)}(\Psi_r^{(m)})\Big)\bigg)\bigg)\\
&\le\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathfrak{E}\bigg(\mathbb{I}_{A_N^{(r,m)}}\beta^{\Theta_r}\bigg(\sum_{l=1}^{\sigma_r(m)}\beta^l\Big(h^{(m)}(\Psi_r^{(m)}+l)-\gamma^{(m)}(\Psi_r^{(m)})\Big)\bigg)\bigg)\\
&\le\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathfrak{E}\bigg(\mathbb{I}_{A_N^{(r,m)}}\beta^{\Theta_r}\mathfrak{E}_{\Psi_r}\bigg(\sum_{l=1}^{\sigma_r(m)}\beta^l\Big(h^{(m)}(\Psi_r^{(m)}+l)-\gamma^{(m)}(\Psi_r^{(m)})\Big)\bigg)\bigg)\\
&=\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathfrak{E}\bigg(\mathbb{I}_{A_N^{(r,m)}}\beta^{\Theta_r}\mathcal{E}^{(m)}\bigg(\sum_{l=1}^{\sigma_r(m)}\beta^l\Big(h^{(m)}(\Psi_r^{(m)}+l)-\gamma^{(m)}(\Psi_r^{(m)})\Big)\bigg|\mathcal{F}^{(m)}_{\Psi_r^{(m)}}\bigg)\bigg)\\
&=\sum_{m=1}^{M}\sum_{r\in\underline{\mathcal{S}}_N}\mathfrak{E}(0) \quad = \quad 0.
\end{aligned}$$

We see that $V(N+1,\sigma,p)\le 0$, and the desired result follows by induction. □

B.3 Step B.3: C-optimality

In the previous subsections, we introduced an allocation problem when the prevailing process is offered as compensation (Definition B.3). We also proved that the optimal value can be achieved by choosing a proper family of allocation time sequences (i.e. σ as in Theorem B.7).

The prevailing reward process Γ for each bandit 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 bandit with the lowest value of Γ first. In particular, we can establish the Robust Gittins index theorem, which we repeat for convenience of the reader.

Theorem B.8 (Theorem 4.3: Robust Gittins theorem)**.** *Suppose that for each $m \in \mathcal{M}$, $(\mathcal{F}_t^{(m)})_{t\geq 0}$ is generated by some underlying process $(\xi_t^{(m)})_{t\geq 1}$. 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

$$\rho_n^* := \min\Big\{m \in \mathcal{M} \quad : \quad m \in \arg\min_k \gamma^{(k)}(\psi_n^{(k)})\Big\}$$

is C-optimal (Definition 3.21) 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 (4.7). 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

$$p_n^* := \min\left\{m \in \mathcal{M} \quad : \quad m \in \arg\min_k \gamma^{(k)}(\Psi_r^{(k)})\right\} \quad \text{on the event} \quad \{\tilde{\eta}_n = r\}.$$

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

$$\begin{aligned}\Big(\{\tilde{\eta}_n = r\} \cap \{p_n^* = m\}\Big) &= \{\tilde{\eta}_n = r\} \cap \bigcap_{k=1}^{m} \Big\{\gamma^{(m)}(\Psi_r^{(m)}) < \gamma^{(k)}(\Psi_r^{(k)})\Big\} \cap \bigcap_{k=1}^{M} \Big\{\gamma^{(k)}(\gamma^{(m)}(\Psi_r^{(m)}) \leq \Psi_r^{(k)})\Big\} \\ &\in \mathcal{F}(\Psi_r).\end{aligned}$$

Hence, p^* is a choice sequence for the allocation sequence σ (Definition 4.8). 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 B.7 and Theorem B.6,

$$\mathfrak{E}\bigg(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1}^*)}(t_n^*) - \Gamma^{(\rho_{n-1}^*)}(t_n^*)\right)\bigg) = 0 \quad \text{for} \quad t_n^* := \sum_{k=0}^{n-1} \mathbb{I}(\rho_k^* = \rho_{n-1}^*).$$

Theorem B.6 also implies that for any allocation strategy (τ, p) (and thus for any simple form ρ),

$$\mathfrak{E}\bigg(\sum_{n=1}^{\infty} \beta^n \left(h^{(\rho_{n-1})}(t_n^\rho) - \Gamma^{(\rho_{n-1})}(t_n^\rho)\right)\bigg) \geq 0.$$

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 $\Gamma^{(m)}(t) = \max_{0\leq\theta\leq t-1} \gamma(\theta)$ and 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 $(\imath, \rho)$. We can write

$$\Gamma^{(\rho_{n-1})}(t_n^\rho) = \sum_{r\in\mathcal{S}_{n-1}} \mathbb{I}(\eta_{n-1} = r)\mathbb{I}(\rho_{n-1} = m)\Gamma^{(m)}(1 + r^{(m)})$$

where $\mathcal{S}_N := \Big\{r \in \mathcal{S} \ : \ \sum_{m=1}^{M} r^{(m)} = N\Big\}$.

Since $\Gamma^{(m)}(1+r^{(m)})$ is $\mathcal{F}^{(m)}_{r^{(m)}}$-measurable, by the Doob–Dynkin lemma, there exists a measurable function $f^{(m)}_r : \mathbb{R}^{r^{(m)}} \to \mathbb{R}$ such that

$$\Gamma^{(m)}(1+r^{(m)}) = f^{(m)}_r(\xi^{(m)}_1, ..., \xi^{(m)}_{r^{(m)}}). \tag{B.3}$$

Note that, while the process $\big(\xi^{(m)}_t\big)$ 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 (B.3) in a similar way to (C.6), we can write $\Gamma^{(\rho_{n-1})}(t^\rho_n)$ as a (measurable) function of $\Big((\eta_k)_{0\le k\le n-1}, (\rho_k)_{0\le k\le n-1}, (\xi^\rho_k)_{0\le k\le n-1}\Big)$. By Definition 4.8 and Remark 4.14, (η_k) and (ρ_k) are adapted to the observed filtration $(\mathcal{H}^\rho_n)_{n\ge 0}$. It follows that $\Gamma^{(\rho_{n-1})}(t^\rho_n)$ is $\mathcal{H}^\rho_{n-1}$-measurable.

Therefore, for each ρ, $C^\rho_0(n) := \Gamma^{(\rho_{n-1})}(t^\rho_n)$ defines a subcompensator at time $N = 0$, and for $\rho = \rho^*$, the cost is fully compensated with $V_0 = 0$.

To construct a 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 C.12). As $\mathcal{F}(r)$ describes the information from all bandits, this needs to be done carefully. Each of our single-bandit filtrations $(\mathcal{F}^{(m)}_t)_{t\ge 0}$ is generated by a discrete-time real-valued process, and $\mathcal{F}(r) = \bigotimes_m \mathcal{F}^{(m)}_{r^{(m)}}$, so 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 $u \le r$. Let $(\rho^{*,\omega_r}_n)_{n\ge N}$ 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

$$\Gamma^{(m)}_{\omega_r}(t) = \max_{r^{(m)}\le\theta\le t-1} \gamma^{(m)}(\theta) \le \max_{u^{(m)}\le\theta\le t-1} \gamma^{(m)}(\theta) = \Gamma^{(m)}_{\omega_u}(t) \quad \text{for all } \ u \le r$$

and are measurable with respect to ω_r. As discussed in Remark 4.5, the optimal strategy $(\rho^{*,\omega_r}_n)_{n\ge N}$ 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, $\Gamma^{(\rho^{*,\omega_r}_{n-1})}_{\omega_r} \le \Gamma^{(\rho^{*,\omega_r}_{n-1})}_{\omega_u}$. We can now unfreeze ω_r and ω_u and, summing over all possible scenarios, show that

$$C^\rho_N(n) := \beta^n \max_{\eta^{(\rho_{n-1})}_N \le\theta\le t^\rho_n - 1} \gamma^{(\rho_n)}(\theta)$$

is decreasing in N. By applying the same argument as earlier, we also have $n \mapsto C^\rho_N(n)$ is $\mathcal{H}^\rho_n$-predictable. Therefore, $(C^\rho_N(n))$ defines a subcompensator for strategy ρ and fully compensates for $\rho = \rho^*$. We set $V_N = 0$ and observe (3.2) is satisfied.

Finally, as $t \mapsto \Gamma^{(m)}_{\omega_r}(t)$ 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

$$\sum_{n=N+1}^{L} C^{\rho^*}_N(n) \le \sum_{n=N+1}^{L} C^\rho_N(n) \qquad \text{for all } \ L \ge N+1. \tag{B.4}$$

In particular, (3.3) is satisfied and therefore ρ^* is C-optimal. □

C Proof of other relevant results

Definition C.1 (Föllmer and Schied [36]). *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* $\operatorname{ess\,inf}$ *if the measurability of the family is obvious.*

Theorem C.2 (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.*

The corresponding results also hold for the $\mathcal{G}$*-essential supremum.*

Proof. See Theorem 1.3.40. in [29] or Theorem A.37 in [36]. □

Theorem C.3 (Proof of Theorem 4.2)**.** *Let* $\gamma(s)$ *be the robust Gittins index (Definition 4.1) (with superscript* (m) *omitted). Then*

$$\gamma(s) = \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t\big|\mathcal{F}_s\big)}$$

where $\mathcal{Q}$ *is a family of probability measure induced in Theorem 3.7.*

Proof. By Corollary A.4 and Theorem 3.7 (together with positive homogeneity), we have that, for any $\tau \in \mathcal{T}(s)$,

$$\operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau} \beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right) \geq 0. \tag{C.1}$$

By Lemma 11.19 (together with the construction of Theorem 11.22) in Föllmer and Schied [36], the family $\big\{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t\big(h(s+t)-\gamma(s)\big)\big|\mathcal{F}_s\big) : \mathbb{Q} \in \mathcal{Q}\big\}$ must be directed upwards. Hence, by Theorem C.2, we can find a family $(\mathbb{Q}_n) \subseteq \mathcal{Q}$ such that

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau} \beta^t\big(h(s+t)-\gamma(s)\big)\right) \xrightarrow[n\to\infty]{} \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau} \beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right). \tag{C.2}$$

Let $\tilde{\Omega}$ be an event with probability one such that the following holds.

1. (C.1) and (C.2) hold.
2. For all $n \in \mathbb{N}$,

$$\frac{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t\big|\mathcal{F}_s\big)} \leq \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t h(s+t)\big|\mathcal{F}_s\big)}{\mathbb{E}^{\mathbb{Q}}\big(\sum_{t=1}^{\tau} \beta^t\big|\mathcal{F}_s\big)},$$

$\mathbb{E}^{\mathbb{Q}_n}\big(\sum_{t=1}^{\tau} \beta^t\big|\mathcal{F}_s\big) \geq \beta$, and

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau} \beta^t\big(h(s+t)-\gamma(s)\big)\middle|\mathcal{F}_s\right) = \mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau} \beta^t h(s+t)\middle|\mathcal{F}_s\right) - \gamma(s)\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau} \beta^t\middle|\mathcal{F}_s\right). \tag{C.3}$$

Fix $\omega \in \tilde{\Omega}$ and $\epsilon > 0$. By (C.1) and (C.2), there exists $n \in \mathbb{N}$ such that

$$\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\big(h(s+t)-\gamma(s)\big)\bigg|\mathcal{F}_s\right)(\omega) \geq -\epsilon.$$

By (C.3), we can rearrange the inequality above and obtain

$$\begin{aligned}\gamma(s)(\omega) &\leq \frac{\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)}(\omega) + \frac{\epsilon}{\mathbb{E}^{\mathbb{Q}_n}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)(\omega)} \\ &\leq \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)}(\omega) + \frac{\epsilon}{\beta}.\end{aligned}$$

As ϵ is arbitrary, it follows that on $\tilde{\Omega}$,

$$\gamma(s) \leq \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)},$$

hence,

$$\gamma(s) \leq \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)}\ \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)}.$$

By Theorem A.10 and Theorem 3.7, we can find $\sigma := \sigma(s,\gamma(s)) \in \mathcal{T}(s)$ such that,

$$\operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t\big(h(s+t)-\gamma(s)\big)\bigg|\mathcal{F}_s\right) = 0,$$

hence, for all $\mathbb{Q} \in \mathcal{Q}$,

$$\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t\big(h(s+t)-\gamma(s)\big)\bigg|\mathcal{F}_s\right) \leq 0.$$

Therefore,

$$\gamma(s) \geq \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t\big|\mathcal{F}_s\right)},$$

and we conclude

$$\gamma(s) \geq \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\sigma}\beta^t\big|\mathcal{F}_s\right)} \geq \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)}\ \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}} \frac{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t h(s+t)\big|\mathcal{F}_s\right)}{\mathbb{E}^{\mathbb{Q}}\left(\sum_{t=1}^{\tau}\beta^t\big|\mathcal{F}_s\right)}.$$

This completes the proof. □

Proposition C.4. *All properties described in Proposition 3.13 hold.*

Proof. (i) is straightforward to prove using the definition directly, as in the case of $\mathcal{E}$.

By Theorem 3.7 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 $\hat{Y}(\omega^{(1)},...,\omega^{(M)}) = X^{(1)}(\omega_1)\times\cdots\times X^{(M)}(\omega_M)$, by Fubini's theorem,

$$\operatorname*{ess\,sup}_{\bigotimes_{m=1}^{M}\mathbb{Q}^{(m)}\in\mathcal{Q}} \mathbb{E}^{\bigotimes_{m=1}^{M}\mathbb{Q}^{(m)}}\left(\prod_{m=1}^{M}X^{(m)}\bigg|\mathcal{F}(S)\right) = \operatorname*{ess\,sup}_{\bigotimes_{m=1}^{M}\mathbb{Q}^{(m)}\in\mathcal{Q}}\ \prod_{m=1}^{M}\mathbb{E}^{\mathbb{Q}^{(m)}}\big(X^{(m)}\big|\mathcal{F}^{(m)}_{S^{(m)}}\big).$$

Then (iii) and (iv) follow by considering different choices of $X^{(m)}$. □

Lemma C.5. *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 $\tau, \sigma \in \mathcal{T}(s)$, 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 C.2. □

Lemma C.6. *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

$$0 \leq f(\tau, Y) - f(\tau, X) \leq L(X - Y) \quad \mathbb{P}\text{-a.s.}$$

(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 $X^* := \operatorname{ess\,inf}\Big\{X \in L^\infty(\mathcal{F}_s) : \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X) \leq 0 \quad \mathbb{P}\text{-a.s.}\Big\}$. 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 $\mathcal{X} := \Big\{X \in L^\infty(\mathcal{F}_s) : \operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X) \leq 0 \quad \mathbb{P}\text{-a.s.}\Big\}$. By (i), $\mathcal{X} \neq \phi$. For a fixed $X \in \mathcal{X}$, by Lemma C.5 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

$$\begin{aligned}\operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f\left(\tau, \min(X, X')\right) &\leq f(\sigma_n, \min(X, X')) = f(\tau_k\mathbb{I}_A + \tau'_k\mathbb{I}_{A^c}, X\mathbb{I}_A + X'\mathbb{I}_{A^c}) \\ &= f(\tau_k, X)\mathbb{I}_A + f(\tau'_k, X')\mathbb{I}_{A^c} \searrow \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X)\mathbb{I}_A + \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X')\mathbb{I}_{A^c} \quad \leq \quad 0.\end{aligned}$$

Hence, $\mathcal{X}$ is downward directed. Therefore, by Theorem C.2, 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

$$\begin{aligned}L(X_n - X^*) &\geq f(\tau_k^n, X^*) - f(\tau_k^n, X_n) \geq \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*) - f(\tau_k^n, X_n) \\ &\nearrow \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*) - \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X_n) \geq \operatorname*{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*).\end{aligned}$$

where the above limit is considered as $k \to \infty$. Now, by taking $n \to \infty$ on the LHS, we obtain $\operatorname{ess\,inf}_{\tau\in\mathcal{T}(s)} f(\tau, X^*) \leq 0$.

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

$$\begin{aligned}
\operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, Y) &\leq f(\sigma_k, Y) = f(\tau_k, X^*)\mathbb{I}_{B^c} + f\Big(\sigma, X^* - \frac{F}{2L}\Big)\mathbb{I}_B \\
&= f(\tau_k, X^*)\mathbb{I}_{B^c} + \bigg(\Big(f\Big(\sigma, X^* - \frac{F}{2L}\Big) - f\left(\sigma, X^*\right)\Big) - F\bigg)\mathbb{I}_B \\
&\leq f(\tau_k, X^*)\mathbb{I}_{B^c} + \left(L\left(X^* - \left(X^* - \frac{F}{2L}\right)\right) - F\right)\mathbb{I}_B \\
&= f(\tau_k, X^*)\mathbb{I}_{B^c} - \frac{F}{2}\mathbb{I}_B \leq f(\tau_k, X^*)\mathbb{I}_{B^c} \quad \searrow \quad \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} f(\tau, X^*)\mathbb{I}_{B^c} \leq 0.
\end{aligned}$$

Hence, $Y \in \mathcal{X}$. By minimality of X^*, it must follow that B is a $\mathbb{P}$-null set. □

Theorem C.7. *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 $\mathcal{F}^s_n := \mathcal{F}_{s+n}$ and define the processes

$$Y_0 := \frac{(C+1) - \lambda}{1 - \beta}, \quad Y_n := \sum_{t=s+1}^{s+n} \beta^t \left(h(t) - \lambda\right) \quad \text{and} \quad Z_n := \operatorname*{ess\,inf}_{\tau \geq n} \mathcal{E}\big(Y_\tau \big| \mathcal{F}^s_n\big)$$

where τ is considered over the space of all stopping times and C is an upper bound given in Assumption 3.1.

Define $\tau^* := \inf\{n \geq 0 : Y_n = Z_n\}$.

By robust representation theorem (Theorem 3.7), we can represent $\mathcal{E}$ as an essential supremum over a familiy of probability measures which satisfy the law of iteration [68, Equation 4]. It then follows Riedel [68, 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. This means that $\tau^* \in \mathcal{T}(s)$ and $Z_0 = \operatorname{ess\,inf}_{\tau \in \mathcal{T}(s)} V_s(\tau, \lambda)$. It remains to prove that τ^* is almost surely finite.

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 C.8. *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*

$$\rho_n = \begin{cases} \rho_{n-1} & \text{if } \sum_{m=1}^{M} \pi_n^{(m)} \geq n, \\ p_{\psi_n} & \text{if } \sum_{m=1}^{M} \pi_n^{(m)} = n - 1, \end{cases} \quad \text{where} \quad \pi_n^{(m)} := \sum_{k=0}^{\hat{F}_n^{(m)} - 1} \tau_k^{(m)}, \tag{C.4}$$

and where

$$\psi_n = \sum_{m=1}^{M} \hat{F}_n^{(m)} \quad \text{and} \quad \hat{F}_n^{(m)} := \min\left\{ f \geq 0 \ : \ \sum_{k=0}^{n-1} \mathbb{I}(\rho_k = m) \leq \sum_{k=0}^{f-1} \tau_k^{(m)} \right\}.$$

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 $\psi_n = \sum_{m=1}^{M} \hat{F}_n^{(m)}$ 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}. $\square$

Definition C.9. *Given a simple form choice sequence ρ, we define the* decision filtration *induced by ρ by*

$$\mathcal{G}_n^\rho := \{A \in \mathcal{F}(T) \;\;:\;\; A \cap \{\eta_n = r\} \in \mathcal{F}(r)\;\} \tag{C.5}$$

where η is the recording sequence corresponding to $(\imath, \rho)$.

Lemma C.10. *The sequence of σ-algebras $(\mathcal{G}_n^\rho)$ given in Definition C.9 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 $m \in \{0\} \cup \mathcal{M}$,

$$A \cap \{\eta_n = r - e^{(m)}\} \in \mathcal{F}(r - e^{(m)}) \subseteq \mathcal{F}(r).$$

Moreover, by definition, $\{\rho_n = m\} \in \mathcal{G}_n^\rho$, it follows that

$$A \cap \{\eta_{n+1} = r\} = \bigcup_{m=0}^{M} \big(A \cap \{\eta_n = r - e^{(m)}\}\big) \cap \big(\{\rho_n = m\} \cap \{\eta_n = r - e^{(m)}\}\big) \in \mathcal{F}(r).$$

Hence, $A \in \mathcal{G}_{n+1}^\rho$. $\square$

Lemma C.11. *With (ξ_n^ρ) as in Remark 4.14 and $(\mathcal{G}_n^\rho)$ in Definition C.9, 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 C.10, ρ_{n-1} is $\mathcal{G}_n^\rho$-measurable. Hence, for a fixed $B \in \mathcal{B}$, we have

$$\{\xi_n^\rho \in B\} \cap \{\eta_n = r\} = \big\{\xi_{r^{(m)}}^{(m)} \in B\big\} \cap \underbrace{\{\eta_n = r\} \cap \{\rho_{n-1} = m\}}_{\in \mathcal{F}(r) \;\text{ as }\; \{\rho_{n-1}=m\} \,\in\, \mathcal{G}_n^\rho} \in \mathcal{F}(r).$$

Therefore, ξ_n^ρ is $\mathcal{G}_n^\rho$-measurable and the last assertion follows from Lemma C.10. $\square$

Theorem C.12. *With $(\mathcal{H}_n^\rho)_{n\geq 0}$ as in Remark 4.14 and $(\mathcal{G}_n^\rho)_{n\geq 0}$ in Definition C.9, $\mathcal{G}_n^\rho = \mathcal{H}_n^\rho$.*

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 C.11, it suffices to show that $\mathcal{G}_n^\rho \subseteq \mathcal{H}_n^\rho$. Recall from Definition 4.8 that for $n \leq L$, the recording sequence η_n takes values in $\mathcal{S}_n := \big\{r \in \mathcal{S} \;:\; \sum_{m=1}^{M} r^{(m)} = n\big\}$.

For a fixed $r \in \mathcal{S}_N$, we define $\mathcal{P}_r$ to be the space of sequences of length $\sum_{m=1}^{M} r^{(m)}$ with values in $\mathcal{M}$ with exactly $r^{(m)}$ replications of $m \in \mathcal{M}$. (e.g. for $\mathcal{M} = \{1, 2\}$, $\mathcal{P}_{(3,1)} = \{(1,1,1,2), (1,1,2,1), (1,2,1,1), (2,1,1,1)\}$).

For $A \in \mathcal{G}_n^\rho$ and $\pi \in \mathcal{P}_r$,

$$A \cap \big\{(\rho_k)_{0\leq k\leq n-1} = \pi\big\} = A \cap \big\{(\rho_k)_{0\leq k\leq n-1} = \pi\big\} \cap \{\eta_n = r\} \in \mathcal{F}(r).$$

By the Doob–Dynkin lemma, there exists a measurable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\begin{aligned} \mathbb{I}_{A\cap\{(\rho_k)_{0\leq k\leq n-1}=\pi\}} &= f\Big(\big(\xi_t^{(m)}\big)_{1\leq t\leq r^{(m)}, m\in\mathcal{M}}\Big) = f\Big(\big(\xi_t^{(m)}\big)_{1\leq t\leq r^{(m)}, m\in\mathcal{M}}\Big)\mathbb{I}_{\{(\rho_k)_{0\leq k\leq n-1}=\pi\}} \\ &= f_\pi\Big(\,\big(\xi_k^\rho\big)_{1\leq k\leq n}\Big)\mathbb{I}_{\{(\rho_k)_{0\leq k\leq n-1}=\pi\}} \end{aligned} \tag{C.6}$$

for some measurable function f_π defined by reordering the input of f by π.

By the inductive hypothesis, $\mathbb{I}_{\{(\rho_k)_{0\leq k\leq n-1}=\pi\}}$ is $\mathcal{G}_{n-1}^\rho$-measurable and thus $\mathcal{H}_{n-1}^\rho$-measurable. Hence, the RHS of (C.6) is $\mathcal{H}_n^\rho$-measurable. Therefore, we know $\mathbb{I}_A = \sum_{r\in\mathcal{S}_N}\sum_{\pi\in\mathcal{P}_r}\mathbb{I}_{A\cap\{(\rho_k)_{0\leq k\leq n-1}=\pi\}}$ is $\mathcal{H}_n^\rho$-measurable, i.e. $A \in \mathcal{H}_n^\rho$, which completes the proof. $\square$

D Numerical algorithm to estimate robust Gittins index

To approximate the value of $\gamma_{k,\beta,T}\,(p, 1/\sqrt{n})$ in the setting of Section 5, we proceed following the rough recipe below. (The code used is available on request.)

Fix a grid of values for $\gamma \in G \subseteq [0, 1]$.

1. Set $V_T^\gamma\Big(p, \frac{1}{\sqrt{n+T}}\Big) = 0$.
2. Assume that we know V_{t+1}^γ. We evaluate the backward recursion
$$V_t^\gamma\Big(p, \frac{1}{\sqrt{n+t}}\Big) = \min\Big\{0, \mathcal{E}_{(t)}\Big((h(t)-\gamma) + \beta V_{t+1}^\gamma\Big(p_{t+1}, \frac{1}{\sqrt{n+(t+1)}}\Big)\Big)\Big|_{(p_t,n_t)=(p,n+t)}\Big\}.$$
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
$$U\Big(\gamma, p, \frac{1}{\sqrt{n}}, T\Big) = \mathcal{E}_{(0)}\Big((h(1)-\gamma) + \beta V_1^\gamma\Big(p_1, \frac{1}{\sqrt{n+1}}\Big)\Big)\Big|_{(p_0,n_0)=(p,n)}.$$
By Snell’s envelope argument,
$$U\Big(\gamma, p, \frac{1}{\sqrt{n}}, T\Big) = \operatorname*{ess\,inf}_{\tau \in \mathcal{T}(s)} \mathcal{E}\Bigg(\sum_{t=1}^{\tau} \beta^t\,(h(t)-\gamma)\,\Bigg|\mathcal{F}_0\Bigg)\Bigg|_{(p_0,n_0)=(p,n)}.$$
4. Repeat step 1-3 for all $\gamma \in G$.
5. Calculate $\gamma_{k,\beta,T}(p, 1/\sqrt{n})$ for a fixed $(p, 1/\sqrt{n})$ by
$$\gamma_{k,\beta,T}\Big(p, \frac{1}{\sqrt{n}}\Big) = \min\Big\{\gamma \in G : U\Big(\gamma, p, \frac{1}{\sqrt{n}}\Big) \le 0\Big\}.$$
6. Repeat the previous step to compute $\gamma_{k,\beta,T}(p, 1/\sqrt{n})$ for other values of $(p, 1/\sqrt{n})$ and obtain the surface by linear interpolation.

References

[1] R. Agrawal, *Sample mean based index policies by $O(\log N)$ regret for the multi-armed bandit problem*, Advances in Applied Probability (1995), 1054–1078. MR1358906

[2] L. An, S. N. Cohen, and S. Ji, *Reflected Backward Stochastic Difference Equations and Optimal Stopping Problems under g-expectation*, arXiv:1305.0887 (2013).

[3] F. Anscombe, *Sequential medical trials*, Journal of the American Statistical Association (1963), 365–383. MR0152096

[4] P. Armitage, *Sequential medical trials*, Blackwell Scientific, 1960. MR0143321

[5] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath, *Coherent measures of risk*, Mathematical Finance (1999), 203–228. MR1850791

[6] 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 (2007), 5–22. MR2303124

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire, *Gambling in a rigged casino: The adversarial multi-armed bandit problem*, Proceedings of IEEE 36th Annual Foundations of Computer Science (1995). MR1619094

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire, *The nonstochastic multiarmed bandit problem*, SIAM Journal on Computing (2003), 48–77. MR1954855

[9] P. Bank and N. El Karoui, *A stochastic representation theorem with applications to optimization and obstacle problems*, The Annals of Probability (2004), 1030–1067. MR2044673

[10] P. Bank and H. Föllmer, *American options, multi–armed bandits, and optimal consumption plans: A unifying view*, Paris-Princeton Lectures on Mathematical Finance 2002 (2002), 1–42. MR2021789

[11] P. Bank and C. Küchler, *On Gittins' index theorem in continuous time*, Stochastic Processes and their Applications (2007), 1357–1371. MR2343944

[12] J. A. Bather, *Recent advances in statistics: Papers in honor of Herman Chernoff on his sixtieth birthday*, ch. Optimal stopping of Brownian motion: a comparison technique, pp. 19–49, Academic Press, 1983.

[13] D. Berry and B. Fristedt, *Bandit problems: Sequential allocation of experiments.*, Chapman & Hall, 1985. MR0813698

[14] T. Bielecki, T. Chen, and I. Cialenco, *Recursive construction of confidence regions*, Electronic Journal of Statistics **11** (2017), no. 2, 4674–4700. MR3724972

[15] J. Bion-Nadal, *Dynamic Risk Measures: Time Consistency and Risk Measures from BMO Martingales*, Finance and Stochastics (2008), 219–244. MR2390189

[16] T. Björk, M. Khapko, and A. Murgoci, *On time-inconsistent stochastic control in continuous time*, Finance Stoch. **21** (2017), 331–360. MR3626618

[17] T. Björk and A. Murgoci., *A theory of Markovian time-inconsistent stochastic control in discrete time*, Finance Stoch. **18** (2014), 545—-592. MR3232016

[18] M. Brezzi and T.L. Lai, *Optimal learning and experimentation in bandit problems*, Journal of Economic Dynamics and Control (2002), 87–108. MR1925627

[19] G. Burtini, J. Loeppky, and R. Lawrence, *A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit*, arXiv:1510.00757v4 (2015).

[20] F. Caro and A. D. Gupta, *Robust control of the multi-armed bandit problem*, Annals of Operations Research (2015), 1–20.

[21] F. Chang and T. L. Lai, *Optimal stopping and dynamic allocation*, Advances in Applied Probability (1987), 829–853. MR0914595

[22] X. Cheng and F. Riedel, *Optimal stopping under ambiguity in continuous time*, Mathematics and Financial Economics (2013), 29–68. MR3023890

[23] S. N. Cohen, *Data-driven nonlinear expectations for statistical uncertainty in decisions*, Electronic Journal of Statistics (2016), 1858–1889. MR3641848

[24] S. N. Cohen, *Uncertainty and filtering of hidden Markov models in discrete time*, Probability, Uncertainty and Quantitative Risk (2017), 1–34. MR4109804

[25] S. N. Cohen, *Data and uncertainty in extreme risks – a nonlinear expectations approach*, Innovations in Insurance, Risk and Asset Management, World Scientific (2018), 135–162.

[26] S. N. Cohen, *Representing filtration consistent nonlinear expectations as g-expectations in general probability spaces*, Stochastic Processes and their Applications (2018), 1601–1626. MR2914765

[27] S. N. Cohen and R. J. Elliott, *A general theory of finite state Backward Stochastic Difference Equations*, Stochastic Processes and their Applications (2010), 442–466. MR2594366

[28] S. N. Cohen and R. J. Elliott, *Backward Stochastic Difference Equations and nearly-time-consistent nonlinear expectations*, SIAM Journal on Control and Optimization (2011), 125–139. MR2765659

[29] S. N. Cohen and R. J. Elliott, *Stochastic calculus and applications*, Birkhäuser, 2015. MR3443368

[30] K. Detlefsen and G. Scandolo, *Conditional and dynamic convex risk measures*, Finance Stochastics (2005), 539–561. MR2212894

[31] N. El Karoui, C. Kapoudjian, E. Pardoux, S. Peng, and M. C. Quenez, *Reflected solutions of Backward SDE's and related obstacle problems for PDE's*, The Annals of Probability (1997), 702–737. MR1434123

[32] 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 (1993), 1232–1236. MR1128498

[33] N. El Karoui and I. Karatzas, *Dynamic allocation problems in continuous time*, The Annals of Applied Probability (1994), 255–286. MR1272729

[34] N. El Karoui, S. Peng, and M. C. Quenez, *Backward Stochastic Differential Equations in finance*, Mathematical Finance (1997), 1–71. MR1434407

[35] H. Föllmer and I. Penner, *Convex risk measures and the dynamics of their penalty functions*, Statistics & Decisions (2006), 61–96. MR2323189

[36] H. Föllmer and A. Schied, *Stochastic Finance: an introduction in discrete time*, De Gruyler, 2016. MR3859905

[37] M. Frittelli and E. R. Gianin, *Putting order in risk measures*, Journal of Banking & Finance (2002), 1473–1486.

[38] M. Frittelli and G. Scandolo, *Risk measures and capital requirements for processes*, Mathematical Finance (2006), 589–612. MR2253119

[39] E. Frostig and G. Weiss, *Four proofs of Gittins’ multiarmed bandit theorem*, Annals of Operations Research (2016), 127–165. MR3509415

[40] J. C. Gittins, *Bandit processes and dynamic allocation indices*, Journal of the Royal Statistical Society (1979), 148–177. MR0547241

[41] J. C. Gittins and D. M. Jones, *A dynamic allocation index for the sequential design of experiments*, Progress in Statistics (Amsterdam: North Holland) (J. Gani, ed.), 1974, pp. 241–266. MR0370964

[42] J. C. Gittins and D. M. Jones, *A dynamic allocation index for the discounted multiarmed bandit problem*, Biometrika **66** (1979), no. 3, 561–565.

[43] Y. Gocgun and A. Ghate, *Lagrangian relaxation and constraint generation for allocation andadvanced scheduling*, Computers and Operations Research (2012). MR2885350

[44] Y. Hu, H. Jin, and X. Y. Zhou, *Time-inconsistent stochastic linear-quadratic control*, SIAM J. Control and Optimization **50** (2012), no. 3, 1548–1572. MR2968066

[45] X. Huo and F. Fu, *Risk-aware multi-armed bandit problem with application to portfolio selection*, Royal Society Open Science (2017). MR3745993

[46] G. N. Iyengar, *Robust dynamic programming*, Mathematics of Operations Research (2005), 257–280. MR2142033

[47] D. Kahneman and A. Tversky, *Prospect theory: Analysis of decision under risk*, Econometrica (1979), 263–292. MR3618580

[48] J. Kay, *Keynes was half right about the facts*, Financial Times (4 August 2015).

[49] F. P. Kelly, *Multi-armed bandits with discount factor near one: The Bernoulli case*, The Annals of Statistics (1981), 987–1001. MR0628754

[50] J. M. Keynes, *A treatise on probability*, Macmillan and Co., 1921, Reprint BN Publishing, 2008. MR3838405

[51] M. J. Kim and A. E.B. Lim, *Robust multiarmed bandit problems*, Management Science (2015), 264–285.

[52] F. H. Knight, *Risk, uncertainty and profit*, Houghton Mifflin, 1921, reprint Dover 2006.

[53] T. C. Koopmans, *Stationary ordinal utility and impatience*, Econometrica (1960), 287–309. MR0127422

[54] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in Applied Mathematics (1985), 4–22. MR0776826

[55] T. Lattimore, *Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits*, Proceedings of Machine Learning Research **49** (2016), 1214–1245.

[56] J. Li, *The k-armed bandit problem with multiple priors*, Journal of Mathematical Economics (2019), 22–38. MR3881745

[57] A. Mandelbaum, *Discrete multi-armed bandits and multi-parameter processes*, Probabability Theory and Related Fields (1986), 129–147. MR0814665

[58] A. Nilim and L. El Ghaoui, *Robust control of Markov decision processes with uncertain transition matrices*, Operations Research (2005), 780–798. MR2171651

[59] S. C. Offwood, *g-Expectations with application to risk measures*, Master's thesis, University of Witwatersrand, 2013.

[60] B. Peleg and M. E. Yaari, *On the Existence of a Consistent Course of Action when Tastes are Changing*, The Review of Economic Studies (1973), 391–401.

[61] S. Peng, *Backward stochastic differential equations, chapter 9: Backward sde and related g-expectation*, Pitman Research Notes in Mathematics, Longman, 1997, 141-159. MR1752680

[62] S. Peng, *Backward stochastic differential equation, nonlinear expectation and their applications*, Proceedings of the international Congress of Mathematics, 2010. MR2827899

[63] S. Peng, *Nonlinear Expectations and Stochastic Calculus under uncertainty*, 2010. MR3970247

[64] R. A. Pollak, *Consistent planning*, The Review of Economic Studies (1968), 201–208.

[65] J. Quiggin, *A theory of anticipated utility*, Journal of Economic Behavior and Organization (1982), 323–343.

[66] J. Quiggin, *Generalized expected utility theory. the rank-dependent model.*, Kluwer Academic, Boston, 1993.

[67] F. Riedel, *Dynamic coherent risk measures*, Stochastic Processes and their Applications (2004), 185–200. MR2073410

[68] F. Riedel, *Optimal Stopping With Multiple Priors*, Econometrica (2009), 857–908. MR2531363

[69] H. Robbins, *Some aspects of the sequential design of experiments.*, Bulletin of the American Mathematical Society (1952), 527–535. MR0050246

[70] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, *A tutorial on thompson sampling*, Foundations and Trends in Machine Learning **11** (2018), no. 1, 1–96. MR3517091

[71] A. Ruszczyński, *Risk-averse dynamic programming for markov decision processes*, Mathematical Programming (2010), 235–261. MR2733564

[72] 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 (2015), 974–980.

[73] R. H. Strotz, *Myopia and inconsistency in dynamic utility maximization*, The Review of Economic Studies (1955 - 1956), 165–180.

[74] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, 1944. MR0011937

[75] R. Weber, *On the Gittins index for multi-armed bandits*, The Annals of Applied Probability (1980), 1024–1033. MR1189430

[76] P. Whittle, *Multi-armed bandits and the Gittins index*, Journal of the Royal Statistical Society: Series B (1980), 143–149. MR0583348

[77] P. Whittle, *Restless bandits: Activity allocation in a changing world*, Journal of Applied Probability (1988), 287–298. MR0974588

[78] Y. Yao, *Some results on the Gittins index for a normal reward process*, Lecture Notes–Monograph Series (2006), 284–294. MR2427855

[79] J. Yong, *Time-inconsistent optimal control problems and the equilibrium HJB equation*, Mathematical Control & Related Fields **2** (2012), no. 3, 271–329. MR2991570

[80] L. Zhou, *A survey on contextual multi-armed bandits*, arXiv:1508.03326 (2016).

Acknowledgments. Samuel Cohen thanks the Oxford-Man Institute for research support and acknowledges the support of The Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1. Tanut Treetanthiploet thanks the University of Oxford for research support while completing this work, and acknowledges the support of the Development and Promotion of Science and Technology Talents Project (DPST) of the Government of Thailand.