Learning In Three Classic Bandit Strategies



Nathaniel V. Korda

St. Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

under the supervision of

Dr. Pierre Tarrès

Michaelmas Term 2011

Acknowledgements

I would like to thank my supervisor, Dr. Pierre Tarrès, for all his guidance in and hard work towards the production of this thesis, Anthony Lee for his collaboration (particularly in the developement of Chapter 4) and stimulating discussions at the office, and both the EPSRC and the Man Group for their financial support. I would also like to thank my many friends in the Stochastic Analysis Group and at the Oxford Man Institute for making my time as a DPhil student a pleasure.

Finally I thank my family for all their love, support and patience over the years.

Abstract

In a Bandit Problem one is repeatedly asked to implement one of a set of actions each of which produce some observable long term behaviour, and one aims to achieve some objective such as maximisation of a reward stream, or consistency. The fundamental dilema one must address is when to choose to explore the behaviour of the actions and when to choose to exploit the information one has gained about them.

In this thesis we provide analyses of three classic Bandit Problem strategies, the Gittins Index Strategy, the Narendra Strategy, and the Thompson Strategy. We are motivated by questions of how reward optimisation in the Bandit Problem setting can be related to learning.

In particular we prove that the Gittins Index Strategy is inconsistent and give a lower bound on rate at which the probability that it does not learn goes to zero as the discount factor tends to one. In doing so we provide new asymptotic explicit upper and lower bounds for the Gittins Indices.

We also provide an analysis of the Narendra Strategy in a quenched context, giving a characterisation of consistency and a study of the rates of convergence that can be observed.

Finally we perform a regret analysis of the Thompson Strategy, showing in particular that for Bernoulli Bandits it achieves logarithmic cummulative regret.

Contents

1 Introduction **4**
Bibliography 18

2 Consistency and Rates of Convergence of the Narendra Strategy for Ergodic Bandits **20**
2.1 Introduction 21
2.2 Preliminaries 27
2.3 Characterisation of Consistency 28
2.4 Proof of Lemmas 2.6, 2.7 and 2.8 33
2.5 Rate of Convergence to 0 38
2.6 Rate of Convergence to 1 39
2.7 Discussion of the Assumptions 50
2.7.1 Assumption (**E1**) 51
2.7.2 Assumption (**E3**) 51
2.7.3 The case $\phi \equiv Id$ 53
2.8 Conclusion 54
Bibliography 56

3 Some Asymptotic Aspects of the Behaviour of the Gittins Index Strategy for Bandits with Bernoulli Payoffs **58**
3.1 Introduction 59
3.2 The Gittins Index and Inconsistency of π_G 65
3.2.1 The Gittins Indices 65
3.2.2 Inconsistency of the Gittins Index Strategy 68
3.3 Explicit Bounds for the Gittins Index 70
3.3.1 Scheme for Deriving the Bounds 70
3.3.2 Tools for Deriving the Bounds 72
3.3.3 The Upper Bound 82

3.3.4 The Lower Bound 83
3.4 A Preliminary Non-Zero Inconsistency Bound 85
3.5 Conclusion 89
Bibliography 90

4 A Consistency and Regret Analysis of the Thompson Strategy for Bandit Problems 92
4.1 Introduction 93
4.2 Consistency 96
4.3 Regret Bounds 98
4.3.1 General Method 98
4.3.2 Uniform Rewards 101
4.3.3 Bernoulli Rewards 102
4.4 Conclusion 110
4.5 Appendix 111
Bibliography 112

Chapter 1

Introduction

Bandit problems comprise a broad class of decision problems that has been widely studied and applied in a number of fields for almost 80 years. Their history goes back to W. R. Thompson in the 1930s whose motivation was found in medical decision making. Through the later half of the 20th century bandit problems recurred in the fields of sequential design of experiments, dynamic programming, stochastic approximation, and computer science, and they have found applications in clinical trials, industrial processes, economics and internet advertising. The broad range of fields treating the subject has led to a number of different notions of the optimality of strategies for this class of problems. In this work we study three classic strategies from this literature, the Gittins Index Strategy, the Narendra Strategy, and the Thompson Strategy, bringing them into new contexts and exploring connections between the different notions of optimality that they were designed to satisfy.

In particular we give a proof of the inconsistency of the Gittins Index Strategy, lower bound the rate at which the probability that this strategy fails to learn goes to zero as the discount factor tends to 1, and give some new explicit upper and lower bounds for the Gittins Indices. We provide an analysis of the Narendra Algorithm in an ergodic context, giving a characterisation of consistency and an analysis of the rates of convergence of the algorithm. Also we provide a study of the original strategy proposed to solve a Bandit problem, the Thompson Strategy, giving a proof of consistency under a mild consistency condition and a regret analysis.

Basic Model We will begin by describing Bernoulli bandits and introduce variations on these as we need them in this introduction.

In a Bernoulli bandit problem an agent repeatedly chooses to employ one of N *actions*, with labels in the set $A = \{1, \ldots, N\}$. We denote the action that the agent chooses to implement at time n by A_n. Associated with each action a is a sequence of i.i.d. Bernoulli random variables $(O_{a,n})_{n\geq 1}$ for which

$$\mathbb{P}(O_{a,n} = 1) = 1 - \mathbb{P}(O_{a,n} = 0) = p_a$$

for some value p_a that is unknown to the agent.

Any initial information given to the agent is denoted by I_0.

If the agent chooses to play action a at time n then he recieves the reward $I_n = O_{a,n}$. Denote by $\mathcal{X}_n$ the (partial) history up to time n of choices of

action and rewards recieved by the agent, and let $\mathcal{H}$ denote the set of all such possible histories:

$$\mathcal{X}_n = (I_0, (A_1, I_1), \ldots, (A_n, I_n)).$$

The agent makes his (random) choice of action at each time n according to some *strategy*, $\pi : \mathcal{H} \to \mathcal{P}(A)$, which is a map from the set of partial histories to the probability space over the set of actions A, and we denote by Π the set of all possible strategies.

The goal of the situation is to find a strategy that the agent can implement so that some functional of the sequence $(I_n)_{n\geq 0}$ has some given property. We call this scenario a *bandit scenario*.

The crucial feature of bandit scenarios is that at each time the agent must choose either to be greedy and exploit the information he currently has about the scenario to obtain immediate reward, or to explore the scenario more by choosing to implement an action which currently promises less than optimal reward in expectation.

Early Work and Consistency Thompson's paper of 1933 [20] was concerned with applying a Bayesian method to the problem of how to repeatedly choose between different possible treatments for a single illness based on past observations of the performance of each individual treatment. As a model for this situation Thompson considered a bandit scenario as described above with $N = 2$, so that the two possible actions represented two possible treatments. He argued that we should follow a strategy (denoted here by π_{Th}) which at time n chooses the action 1 according to the posterior probability that action 1 is better than action 2.

If $\hat{p}_{a,n}$ denotes a random variable with the distribution equal to the posterior distribution for the unknown success probability p_a derived from the observations $\mathcal{X}_n$, then we can define this strategy by:

$$\pi_{Th}(\mathcal{X}_n)(1) := \mathbb{P}(\hat{p}_{1,n} > \hat{p}_{2,n} \mid \mathcal{X}_n).$$

Although Thompson's problem and strategy are interesting his paper did not set out a good framework for generalising the method and the subject lay dormant for almost 20 years.

However with the advent of the theory of sequential design of experiments in the 1950s authors such as Robbins and Monro [17, 16] reformulated the

problem as one of learning. Their approach with regard to Bandit scenarios is encapsulated in the question of satisfying the so-called "Average Reward Criterion", thus named in the later theory of Markov Decision Processes:

Problem 1.1 (Consistency) *Can we find a strategy π for which*

$$\mathbb{P}_\pi\left(\frac{1}{n}\sum_{i=0}^{n-1} I_{i+1} \to \max_{a\in A} p_a \text{ as } n\to\infty\right) = 1? \tag{L}$$

A strategy is said to be *consistent* if it satisfies (L) and *inconsistent* otherwise. Indeed we give a proof that the Thompson Strategy, π_{Th}, is consistent under mild model consistency condition (see Chapter 4), and there are many other consistent strategies in the literature.

Consistency is often taken as a weak definition of learning.

The Narendra Algorithm In the closely related field of Learning Automata we find another early strategy, the Narendra Algorithm (denoted here by π_{Na}), originally proposed by Narendra and Shapiro in 1969 [15]. In the case when $N = 2$ the strategy is defined by the evolution of the random variable $X_n := \pi_{Na}(\mathcal{X}_n)(1)$:

$$X_0 = x \in [0,1],$$
$$X_{n+1} = \begin{cases} X_n + \gamma_n(1 - X_n) & \text{if at time } n \text{ we employ action 1 and win,} \\ X_n - \gamma_n X_n & \text{if at time } n \text{ we employ action 2 and win,} \\ X_n & \text{otherwise,} \end{cases}$$

where the step sequence $(\gamma_n)_{n\geq 0} \in (0,1)^{\mathbb{N}}$ is a deterministic sequence satisfying $\sum_{k=0}^{\infty}\gamma_k = \infty$ and is chosen a priori along with x.

Consistency is a key type of optimality demanded in the learning automata literature and the Narendra Algorithm was originally reported to be consistent. However much later work in the 1990s and 2000s by Lamberton, Pagès, and Tarrès [[18]; [13]] has shown that, while X_n does indeed converge to a $\{0,1\}$-valued random variable under no extra assumptions on the step sequence [13], the consistency of the strategy requires that $\gamma_n \to 0$ and at a sufficiently fast rate, depending on the values of p_1 and p_2.

In [12] Lamberton and Pagès give a characterisation of consistency. More interestingly they also give a detailed analysis of the rates of convergence of

the strategy (i.e. the rate of convergence to 0 or 1 of the random variable X_n, as $n \to \infty$). Considering the sequences $(\gamma_n)_{n\geq 0}$ with the form

$$\gamma_n = \left(\frac{C}{C+n}\right)^{\alpha} \text{ where } \alpha \in (0,1] \text{ and } C > 0,$$

we reproduce below a summary of their results printed in [12] for the case when $p_1 > p_2$:

- the strategy is inconsistent if

$$(\alpha \in (0,1)) \text{ or if } \left(\alpha = 1 \text{ and } C > \frac{1}{p_2}\right);$$

- the strategy is consistent if

$$\alpha = 1 \text{ and } C \leq \frac{1}{p_2};$$

- if $\alpha = 1$ and $C \geq \frac{1}{(p_1 - p_2)}$, then the rate of convergence on $\{X_n \to 1\}$ is

$$n^{-Cp_1};$$

- if $\alpha = 1$ and $\frac{1}{p_1} < C < \frac{1}{(p_1 - p_2)}$, then *two* rates of convergence occur with positive probability on $\{X_n \to 1\}$,

$$n^{-Cp_1} \text{ and } n^{-C(p_1 - p_2)};$$

- if $\alpha = 1$ and $C \leq \frac{1}{p_1}$ then the rate of convergence on $\{X_n \to 1\}$ is

$$n^{-C(p_1 - p_2)}.$$

From a practical point of view this analysis implies that if the agent is given no initial information about the values of p_1 and p_2, then the only practical choice of step sequence is $\gamma_n = \frac{1}{1+n}$, yielding only a slow rate of convergence.

Ergodic Bandits In [19] Tarrès and Vanderkerkhove investigated the performance of the Narendra Algorithm in the context of an Ergodic Bandit scenario. In an Ergodic Bandit the agent does not receive rewards according to a sequence of observed Bernoulli random variables. Instead there are predefined, deterministic, but unknown sequences $(\eta_{a,n})_{n\geq 0} \in \{0,1\}^{\mathbb{N}_0}$ for each action a, and if at time n the agent chooses to employ action a he receives a reward of $\eta_{a,n}$. The sequences are assumed to be ergodic, in the sense that their Cesaro sums converge to some fixed number:

$$\forall a \in A,\ \frac{\sum_{k=0}^{n} \eta_{a,k}}{n} \to p_a \in (0,1)\,.$$

This generalisation of a bandit scenario is motivated by wanting to consider problems of learning in which we do not wish to assume an underlying probabilistic structure in the environment. Proving consistency in this context provides a means of demonstrating robustness of strategies to different fields of application.

We describe the strength of the ergodicity in the sequences $(\eta_{a,n})_{n\geq 0}$ by an assumption of the form:

(E1ϕ) ϕ is a non-decreasing, concave function with the property that

$$\frac{\max_{a\in A}\left\{\sum_{k=1}^{n}\left(\eta_{a,k}-p_a\right)\right\}}{\phi\left(n\right)} \to 0.$$

Note that in the weakest case we may take $\phi \equiv Id$ and recover the assumption that the Cesaro sums converge. To produce a comparison with the case of i.i.d. Bernoulli payoffs from the actions we can appeal to the Law of the Iterated Logarithm to provide a suitable form of ϕ.

Tarrès and Vanderkerkhove have shown that the behaviour of the Narendra Algorithm in this ergodic context depends not only on the choice of the sequence $(\gamma_n)_{n\geq 0}$, but also on the forms of ϕ that the sequences $(\eta_{a,n})_{n\geq 0}$ admit. In particular they show that under the assumption

(E1) **(E1ϕ)** holds for some function ϕ which satisfies

$$\sup_{n\to\infty} \gamma_n \phi\left(n\right) < \infty$$

X_n converges to a $\{0,1\}$-valued random variable, which we denote by X_∞, and further more demonstrate sufficient conditions for consistency [19].

In this thesis we extend the analysis in [19] and prove a quenched version of the annealed result achieved by Lamberton and Pagès in [12]:

Theorem 1.2 *Suppose $p_1 > p_2$, that*

$$\liminf_{n\to\infty} \frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} > -\pi,$$

*and that **(E1)** holds.*

1. *Suppose that $\Phi^a_{n,\gamma} := \sum_{k=1}^n (p_a - \eta_{a,n})\gamma_k$ converges for $a = 1, 2$. Then for $x \in (0,1)$,*

$$\mathbb{P}_x(X_\infty = 0) > 0 \iff \sum_{n\geq 1}\prod_{k=1}^n (1 - \eta_{2,k}\gamma_k) < \infty^1$$

2. *If we assume $\sum_{k\geq 1}^\infty \gamma_k^2 < \infty$ and that $\Phi^2_{n,\gamma}$ converges, then*

$$\sum_{n\geq 1}\prod_{k=1}^n (1 - p_2\gamma_k) < \infty \iff \sum_{n\geq 1}\prod_{k=1}^n (1 - \eta_{2,k}\gamma_k) < \infty.$$

We can also generalise the results on the rates of convergence given in [12]. Under the assumptions that **(E1ϕ)** holds for some function ϕ, such that $\sup_{n\to\infty} \gamma_n\phi(n) < \infty$, and that $\Phi^a_{n,\gamma}$converges for $a = 1, 2$ we can recover the same summary of results contained in [12] and printed above with only one change to the conditions for inconsistency:

Let $(\gamma_n)_{n\geq 0}$ have the form $\gamma_n = \left(\frac{C}{C+n}\right)^\alpha$ and $p_1 > p_2$,

- then the strategy is inconsistent if

$$\left(\alpha \in \left(\frac{1}{2}, 1\right)\right) \text{ or if } \left(\alpha = 1 \text{ and } C > \frac{1}{p_2}\right).$$

To understand the assumptions involving ϕ better, if $\gamma_n = \frac{C}{C+n}$, $C > 0$, it can be shown that if we can take $\phi(n) = n(\log n)^{-(1+\epsilon)}$ then $\Phi^a_{n,\gamma}$ converges for $a = 1, 2$ and $\lim_{n\to\infty} \gamma_n\phi(n) = 0$. A more detailed discussion of these assumptions is given in Chapter 2.

[1] When analysing the Narendra Algorithm, for any event A and $x \in (0,1)$ we use the notations $\mathbb{P}_x(A) := \mathbb{P}(A \mid X_0 = x)$ and $\mathbb{E}_x(A) := \mathbb{E}(A \mid X_0 = x)$ throughout the thesis.

Reward Optimisation Strategies Another strand of research in Bandit problems grew out of the theory of Dynamic Programming developed by Bellman, Blackwell and others in the 1950s and 1960s (see [1] and [3]). In this context the Bandit scenario recurred as a special case of more general dynamic programming problems in which reward optimisation is the goal:

We assume that the initial information $\mathcal{X}_0 = I$ provides us with, for each action $a \in A$, a prior distribution, $\theta_{a,0}$, for the expected reward generated by employing action a at any time. The payoffs of the bandit are assumed to be Bernoulli with the success probability at time $n \geq 1$ of action a being drawn from the posterior distribution $\theta_{a,n-1}$ for the expected reward generated by employing action a given partial history $\mathcal{X}_{n-1}$ and that the payoffs of the actions are independent of each other.

Under this assumption we suppose further that the rewards, $(I_n)_{n\in\mathbb{N}}$, are discounted by some summable sequence of real numbers $\mathcal{A} = (\alpha_n)_{n\geq 0}$, which we shall call a *discount sequence*. Then if the agent employs a particular strategy π, and is given initial information I, we define his *total expected reward* to be

$$V_\pi(I) = \mathbb{E}_\pi\left(\sum_{n=0}^{\infty} \alpha_n I_{n+1} \mid \mathcal{X}_0 = I\right).$$

If a strategy π^* exists for which

$$V_{\pi^*}(I) = V(I) = \sup_{\pi\in\Pi} V_\pi(I)$$

we say that π^* is $\mathcal{A}$-*optimal*. This leads us to our dynamic programming problem:

Problem 1.3 (Reward Optimisation) *For a given discount sequence $\mathcal{A}$ can we find an $\mathcal{A}$-optimal strategy?*

If we assume the discount sequence $\mathcal{A}$ has the form

$$\alpha_n = \begin{cases} 1 & \text{for } n \leq T, \\ 0 & \text{for } n > T, \end{cases}$$

for some fixed, finite time T, then an $\mathcal{A}$-optimal strategy may be simply computed using Bellman's Optimality Equation and a backwards induction iteration (although for large T this is computationally intensive). In fact it is true that for any discount sequence $\mathcal{A}$ there exists an $\mathcal{A}$-optimal strategy

(a proof can be found in the book [2] by Berry and Fristedt), although for most discount sequences with infinitely many non-zero elements, an optimal strategy is intractable to specify.

The Gittins Index Strategy In the special case of geometric discount sequences (i.e. when $\mathcal{A} = (\alpha^n)_{n\geq 0}$ for $\alpha \in (0,1)$) Gittins and Jones demonstrated an *index strategy* that is $\mathcal{A}$-optimal in 1974 [6]. In an index strategy first the agent computes a table of values, or indices,

$$\iota : \mathbb{Z}_+^2 \to \mathbb{R}$$

so that if, at time n, he has observed $s_{a,n}$ successes and $f_{a,n}$ failures from action a he can assign to that action the value, or index, $\iota(s_{a,n}, f_{a,n})$. The strategy then chooses an action with current maximal index:

$$\pi(\mathcal{X}_n)\left(\arg\max_{a\in A}\{\iota[(s_{a,n}, f_{a,n})]\}\right) := 1.$$

The index that Gittins and Jones described is now called the Gittins Index, and we denote it by ν.

In the simple Bandit scenario with independent Bernoulli rewards the values of the Gittins Index itself, however, are only given implicitly in the literature. In order to implement the Gittins Index Strategy one needs to numerically approximate these solutions, and if one wants to play for arbitrarily long, one must approximate the indices to an arbitrary degree of accuracy. Furthermore Berry and Fristedt proved in 1982 that Index strategies cannot be optimal for any other form of discount sequence [2].

Despite these difficulties Index strategies have been generalised in many directions, with notable work through the 80s, 90s, and 2000s on continuous time bandit scenarios by Mandelbaum, El Karoui and Karatzas [8, 9] among other authors.

However in this thesis we ask a different question of the Gittins Index Strategy, namely whether or not it is consistent, and if not, just how inconsistent it is. In [10] Frank Kelly provided an analysis of the asymptotic behaviour of the Gittins Index Strategy using implicit upper and lower bounds for the function ν. In particular he showed that up to some stopping time, T_{LF}, which almost surely becomes infinite as $\alpha \to 1$, the Gittins Index Strategy strategy follows the Strategy of Least Failures. Under the Strategy of

Least Failures the agent first finds the set of actions with the current minimal amount of failures observed and chooses to implement an action at random from these:

$$\pi_{LF}\left(\mathcal{X}_n\right)\left(\arg\min_{a\in A}\left\{f_{a,n}\right\}\right) := 1.$$

Kelly further showed that, almost surely, after some random time T_R the Gittins Index Strategy will only play one button for ever. Since the Strategy of Least Failures will play every button infinitely often with probability 1, $T_R > T_{LF}$ almost surely.

Implicit in these results is the fact that the Gittins Index Strategy is inconsistent for every value of α. We give an elementary proof of this using the explicit lower bound:

$$\nu\left(s,f\right) \geq \frac{s+1}{s+f+2}.$$

However without an explicit upper bound, something lacking in the literature, it is difficult to lower bound the probability that the Gittins Index Strategy chooses to play the wrong button for ever. In joint work with Pierre Tarrès we address this problem and give new, explicit upper and lower bounds for the Gittins Indices for large values of $s+f$:

Theorem 1.4 (Explicit Asymptotic Bounds for the Gittins Index) *There is an asymptotic upper bound for the Gittins Index $\nu_U : \mathbb{Z}_+^2 \to \mathbb{R}$ given by*

$$\nu_U\left(s,f\right) = x + \frac{\alpha}{1-\alpha}x\left(1-x\right)t$$

where

$$x = \frac{s+1}{s+f+2} \text{ and } t = \frac{1}{s+f+2},$$

and it is valid when $t < \frac{1-\alpha}{\alpha}$.

There is an asymptotic lower bound for the Gittins Index $\nu_L : \mathbb{Z}_+^2 \to \mathbb{R}$ given by

$$\nu_L\left(s,f\right) = x + \frac{\alpha}{1-\alpha x}x\left(1-x\right)t + O\left(t^2\right).$$

and it is valid for sufficiently small t.

Using the new explicit upper bound we are able to derive a preliminary lower bound for the rate at which the probability that π_G fails to learn goes to zero as $\alpha \to 1$ in the case when there are just two actions available, and when the parameter of the worse action is less than $\frac{1}{2}$:

Proposition 1.5 *Assume that* $|A| = 2$*, that* $p_2 \leq \frac{1}{2}$ *and that* $p_1 > p_2$*. If*

$$E := \{A_n = 1 \text{ finitely often}\}$$

then for any fixed $\epsilon > 0$

$$\lim_{\alpha \to 1} \mathbb{P}(E) \cdot (1 - p_1)^{\frac{\alpha}{1-\alpha}} \geq C(p_2, \epsilon) > 0$$

where $C(p_2, \epsilon)$ *is a constant for fixed* ϵ*.*

We know from [10] that $\mathbb{P}(E) \to 0$ as $\alpha \to 1$. However from the same paper we know that π_G itself is tending towards the strategy of least failures as $\alpha \to 1$, which, after a brief consideration, is easily shown to be inconsistent. It seems that learning in the sense of consistency is not important for total expected reward optimisation when $\mathcal{A}$ is geometric.

Outside the context of learning the explicit bounds for the indices almost give the second term in the Taylor expansion of the Gittins Index around the time $t = 0$. If the bounds could be improved to give this second term they could potentially be used to approximate the Gittins Index more efficiently, for example by providing a more accurate initial condition to begin the iterative calculation normally employed to calculate the indices [7].

Regret Analysis A third major strand of research on Bandit Scenarios has been that concerned with finding strategies which minimise some form of *regret*.

The *cumulative regret* of a strategy up to a finite time horizon n is defined to be:

$$\mathcal{R}_\pi(n) = \mathbb{E}_\pi \left(\sum_{k=1}^{n} (p_{\max} - I_k) \right)$$

where

$$p_{\max} = \max_{a \in A} \{p_a\}.$$

The first paper to introduce the notion of cumulative regret was [11] by Lai and Robbins in 1985, and many strategies designed to minise cumulative regret have been proposed since (see for example [5]). Lai and Robbins proved that under a strategy π with the property that, for any values of $\{p_a\}_{a \in A}$,

$$\mathcal{R}_\pi(n) = o(n^\alpha) \text{ for all } \alpha \in (0, 1], \text{ almost surely,} \tag{1.1}$$

the cumulative regret satisfies

$$\liminf_{n\to\infty} \frac{\mathcal{R}_\pi(n)}{\log n} \geq \sum_{p_j \neq p_{\max}} \frac{p_{\max} - p_j}{I\left(p_j || p_{\max}\right)} \tag{1.2}$$

where $I\left(p_j||p_{\max}\right)$ is the Kullback-Leibler divergence and is just given by

$$I\left(p_j||p_{\max}\right) = p_{\max} \log \frac{p_{\max}}{p_i} + (1 - p_{\max}) \log \frac{(1 - p_{\max})}{(1 - p_i)}.$$

Observe that

$$\begin{aligned}
\mathcal{R}_\pi(n) &= \mathbb{E}_\pi \left(\sum_{k=1}^{n} \left(p_{\max} - r\left(I_k\right) \right) \right) \\
&= \mathbb{E}_\pi \left(\sum_{a=1}^{N} \left(p_{\max} - p_a \right) t_{a,n} \right)
\end{aligned}$$

where

$$t_{a,n} := \sum_{k=1}^{n} \mathbf{1}_{\{A_n = a\}}, \ a = 1, 2.$$

Hence it follows that any strategy satisfying (1.1) for any $\{p_a\}_{a\in A}$ must be consistent.

Furthermore Lai and Robbins demonstrated a strategy obtaining the rate (1.2), thus proving that this asymptotic lower bound on the cumulative regret is tight.

In this thesis we are able to show the following result on the cummulative regret of the Thompson Strategy, π_{Th}:

Theorem 1.6 *Assume that $N = 2$. Then $R_{\pi_{Th}}(n) = O(\ln n)$.*

Numerical experiments, not included in the thesis, show that the Thompson Strategy performs better than state-of-the-art regret-minimising algorithms such as the KL-UCB algorithm.

Discussion of Notions of Optimality In the case when $N = 2$ and $p_1 > p_2$ the question of asymptotic bounds for the regret of a consistent strategy π seems, at first glance, closely connected to the question of the rate of convergence to zero of the quantity $\pi(\mathcal{X}_n)(2)$, which we considered for the Narendra Algorithm.

Suppose we know that when $N = 2$ and $p_1 > p_2$

$$\lim_{n\to\infty} \frac{f(n)}{\pi(\mathcal{X}_n)(2)} = O(1) \tag{1.3}$$

for some decreasing positive function f. Then since, by the Strong Law of Large Numbers for martingales (see for example [21], page 123)

$$\lim_{n\to\infty} \frac{t_{2,n}}{\sum_{k=1}^{n} \pi(\mathcal{X}_n)(2)} = 1 \text{ almost surely}$$

we might think to compare $t_{2,n}$ with $\sum_{k=1}^{n} f(n)$.

However consider the special case of Epsilon-Decreasing Strategy [4] which is defined as follows: Let $\bar{p}_{a,n}$ be the empirical mean reward from action a at time n, and take a deterministic sequence $(\epsilon_n)_{n\geq 0} \in (0,1)^{\mathbb{N}}$ for which $\epsilon_n \to 0$ as $n \to \infty$. Then we define the Epsilon-Decreasing Strategy by setting

$$\pi_{(\epsilon)}(\mathcal{X}_n)(1) := \begin{cases} 1-\epsilon_n & \text{when } \bar{p}_{1,n} > \bar{p}_{2,n}, \\ \epsilon_n & \text{otherwise.} \end{cases}$$

Suppose we choose the sequence $(\epsilon_n)_{n\geq 0}$ so that $\sum_{n\geq 0} \epsilon_n = \infty$, then it is an easy consequence of the second Borel Cantelli Lemma that the strategy $\pi_{(\epsilon)}$ plays both arms infinitely often, and thus is consistent. Now since both actions will be implemented infinitely often we also know that $\{\bar{p}_{1,n} < \bar{p}_{2,n}\}$ occurs for only finitely many n, almost surely, and hence

$$\lim_{n\to\infty} \frac{\epsilon_n}{\pi(\mathcal{X}_n)(2)} = 1 \text{ almost surely.}$$

However

$$\sum_{k=1}^{n} \pi(\mathcal{X}_k)(2) = \sum_{k=1}^{n\wedge T} (1-\epsilon_k) + \sum_{k=n\wedge T+1}^{n} \epsilon_k$$

where $T := \sup\{n \geq 1 : \bar{p}_{1,n} < \bar{p}_{2,n}\}$. Hence in order to make a statement about the growth of $t_{2,n}$ it is clearly also nessecary to have some knowledge of

the distribution of T, about which a statement like (1.3) would have nothing to say.

In fact it was shown by Auer, Cesa-Bianchi and Fischer [4] that the Epsilon Decreasing Strategy can achieve logarithmic regret asymptotically, although they do not find that the constant given in (1.2) can be achieved.

Minimizing the expected regret at a finite time n when the Bandit problem itself is drawn from a set of prior distributions for the success parameters $\theta = \{\theta_{a,0}\}_{a \in A}$ is exactly the same problem as maximizing the total expected reward for a discount sequence of the form

$$\alpha_i = \begin{cases} 1 & \text{if } i \leq n, \\ 0 & \text{if } i > n. \end{cases}$$

However in this context of a known finite time horizon, there is no reason for an $\mathcal{A}$-optimal strategy to be a learning strategy in the sense of consistency. Furthermore we demonstrate in Chapter 3 that when the discount sequence $\mathcal{A}$ is geometric there is still no nessecity for the strategy to be consistent, even though the strategy will recieve rewards for ever. The question of whether there exist discount sequences $\mathcal{A}$ such that $\mathcal{A}$-optimal strategies must be consistent remains open.

As we mentioned earlier, consistency, especially in early literature, is often taken to be the basic meaning of learning in a bandit scenario. Regret minimisation offers an alternative, more refined notion of learning together with a natural means of interpreting the "rate of learning". There are other notions of learning in the literature which we unfortunately do not address in this thesis, such as "identification of the best arm" [14], for which a strategy attempts to minimise the number of steps needed to identify a near-optimal action with high probability.

The thesis is structured as follows. In Chapter 2 we study the generalisation of the Narendra Algorithm to the ergodic case. Chapter 3 addresses learning properties of the Gittins Indices and gives the proofs of the new asymptotic upper and lower bounds. Chapter 4 revisits the original algorithm proposed by Thompson, proving its consistency and providing a regret analysis.

Bibliography

[1] Bellman. A problem of sequential design of exeriments. *Sankhya: The Indian Journal of Statistics*, 16:221 – 229, 1956.

[2] D. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

[3] D. Blackwell. Discounted dynamic programming. *The Annals of mathematical Statistics*, 36:226 – 235, 1965.

[4] N. Cesa-Bianchi, P. Fischer, and P. Auer. Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[5] Aurlien Garvier and Olivier Capp. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT, Hungary*, 2011.

[6] John Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 2:148–177, 1979.

[7] John Gittins. *Multi-Armed Bandit Allocation Indices*. Wiley, 1989.

[8] N. El Karoui and I. Karatzas. Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, 2:255 – 286, 1994.

[9] H. Kaspi and A. Mandelbaum. Multi-armed bandits in descrete and continuous time. *The Annals of Applied Probability*, 8:1270–1290, 1998.

[10] F. Kelly. Multi-armed bandits with discount factor near one: The bernoulli case. *The Annals of Statistics*, 9:987–1001, 1981.

[11] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4 – 22, 1985.

[12] D. Lamberton and G. Pagès. How fast is the bandit. *Stochastic Analysis and Applications*, 26:603–623, 2008.

[13] D. Lamberton, G. Pagès, and P. Tarrès. When can the two-armed bandit algorithm be trusted. *Annals of Applied Probability*, 14:1424–1454, 2004.

[14] Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623 – 648, 2004.

[15] K. S. Narendra and I.J. Shapiro. Use of stochastic automata for parameter self-optimization with multimodal performance criteria. *IEEE Transactions on Systems Science and Cybernetics*, 5:352 – 360, 1969.

[16] H. Robbins. Some aspects of the sequential design of experiments. *The Annals of Mathematical Statistics*, 30:527 – 535, 1952.

[17] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400 – 407, 1951.

[18] P. Tarrès. *Algorithmes Stochastiques et Marches Aleatoires Renforcees*. PhD thesis, ENS Cachan (France), 2001.

[19] P. Tarrès and P. Vandekerkhove. On ergodic two-armed bandits. Accepted to The Annals of Applied Probability (to appear 2011).

[20] W.R. Thompson. On the likelihood that one unkown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[21] D. Williams. *Probability with Martingales*. CUP, 1991.

Chapter 2

Consistency and Rates of Convergence of the Narendra Strategy for Ergodic Bandits

Joint work with Pierre Tarrès

Abstract

In recent work [7] Tarrès and Vandekerkhove studied the convergence properties of the Narendra Strategy, a stochastic approximation procedure, for the 2-Armed Bandit under weak ergodic assumptions on the payoff sequences (converging Cesaro sums). We bring techniques from this paper together with arguments from [2] to provide a quenched version of the results from the latter. In particular we prove a characterisation of consistency beyond that of Tarrès and Vandekerkhove and also give an analysis of the rates of convergence of the strategy. We find that the behaviour of the strategy is robust to generalising from annealed to quenched contexts.

2.1 Introduction

An agent is repeatedly presented with a choice of N actions $A = \{1, \ldots, N\}$. At each time n if the agent chooses to employ action a he obtains a payoff of $\eta_{a,n} \in \{0, 1\}$. Assuming that the payoffs from the actions exhibit some long term behaviour (for example converging Cesaro sums), the agent's aim is to ascertain which of the actions has the most lucrative potential without spending too much time employing the worse actions. Problems of this type are referred to in the literature as N-Armed Bandit problems in analogy with the problem of how a gambler should choose between N slot machines in a casino that he frequents.

In recent work Tarrès and Vandekerkhove studied the behaviour of the Narendra Algorithm, a stochastic approximation procedure for this problem, under weak convergence assumptions on the Cesaro sums of the payoff sequences $(\eta_{a,n})_{n\geq 1}$, $a \in A$. Despite its early appearance in the literature and its apparent simplicity the Narendra Algorithm has only recently been well analysed. In [4] Lamberton, Pagès and Tarrès showed that for i.i.d. Bernoulli payoffs the algorithm can be either consistent or inconsistent, depending on the choice of the step size parameter. Further work by Lamberton and Pagès [2] showed that, again for i.i.d. Bernoulli payoffs, the rate of convergence of the algorithm is not governed by a Central Limit Theorem, unlike most stochastic approximation algorithms.

Tarrès and Vandekerkhove relaxed the assumptions on the payoff sequences and provided a quenched generalisation of annealed results in [7].

In particular they prove conditions under which the Narendra Algorithm is consistent, and they give a sufficient condition under which the procedure is inconsistent.

In this work we bring the methods developed by Tarrès and Vandekherkove together with arguments introduced by Lamberton and Pagès in [2] to analyse the speed of convergence of the Narendra Algorithm in a quenched setting, and to give a characterisation of consistency beyond that of Tarrès and Vanderkerkhove. We thus achieve a generalisation of results in [2]. Our motivations are both theoretical and practical; we further demonstrate the general applicability of the methods in [7] to analyse deterministic payoff sequences with converging Cesaro sums; we also show that the behaviour of the Narendra Algorithm is robust to annealed, and thus non-stationary, as well as quenched contexts.

Model and Algorithm For simplicity of notation we restrict ourselves to describing the 2-Armed Bandit, i.e. a bandit problem with two possible actions.

Let $(U_n)_{n\geq 1}$ be a sequence of uniform, independent random variables on $[0,1]$, and let $(\gamma_n)_{n\geq 1} \in (0,1)^{\mathbb{N}}$ be a deterministic sequence for which $\Gamma_n := \sum_{1\leq k\leq n} \gamma_k \to \infty$. We define the Narendra Algorithm by giving the evolution of X_n, the probability that the agent chooses to employ action 1 at time n, in terms of the sequences $(U_n)_{n\geq 1}$, $(\gamma_n)_{n\geq 1}$ and $(\eta_{a,n})_{n\geq 1}$, defining

$$X_0 := x \in (0,1)\,,$$

$$\text{and } X_{n+1} = \begin{cases} X_n + \gamma_{n+1}(1-X_n) & \text{on } \{U_{n+1} \leq X_n\} \cap \{\eta_{1,n+1} = 1\}\,, \\ X_n - \gamma_{n+1} X_n & \text{on } \{U_{n+1} > X_n\} \cap \{\eta_{2,n+1} = 1\}\,, \\ X_n & \text{otherwise.} \end{cases} \tag{2.1}$$

The algorithm thus changes its behaviour at the next time step by naively increasing the probability of implementing an action if it gains a reward from that action.

In this work we assume, as in [7], that the payoff sequences $(\eta_{a,n})_{n\geq 1}$ are deterministic but unknown, and we impose an averaging condition:

$$\frac{\sum_{n\geq k\geq 1} \eta_{a,k}}{n} \to p_a \in (0,1)\,,\ a = 1,2. \tag{E}$$

Let $R_n := \max_{a=1,2} \left\{ \sum_{n \geq k \geq 1} (\eta_{a,k} - p_a) \right\}$ and, for any function $\phi : \mathbb{N} \to \mathbb{R}_+$, let (E$\phi$) denote the condition that $R_n / \phi(n) \to 0$. Notice that taking $\phi(n) := n$ in the condition (Eϕ) we recover the averaging condition (E).

We view condition (E) as a relaxation of the assumption in [2] and [4] that the payoff sequences are stationary. However the condition retains the assumption that there is some averaging behaviour to learn from the sequences. Accordingly we find in this chapter conditions involving the nature of this averaging behaviour under which we can generalise the behaviour of the Narendra Algorithm as explicited in [2].

We will work with various combinations of the following assumptions on the step sequences:

(S1) $(\gamma_n)_{n \geq 1}$ is a non-increasing sequence;

(S2) $\gamma_n = O\left(\Gamma_n e^{-p_2 \Gamma_n}\right)$;

(S3) $\liminf_{n \to \infty} \left(\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} \right) > -\pi := -(p_1 - p_2)$;

(S3') $\liminf_{n \to \infty} \left(\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} \right) > \pi$;

(S4) $\sum_{n \geq 1} \gamma_n \left(\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - \pi \right)^+ < \infty$;

(S5) $\sum_{n \geq 1} \gamma_n^2 < \infty$;

Note that (**S3'**) $\Longrightarrow$ (**S1**) $\Longrightarrow$ (**S3**), (**S3'**) contradicts (**S4**), and (**S2**) $\Longrightarrow$ (**S5**) (this last implication is shown in [4]).

Previous work on the case when the sequences $(\eta_{a,n})_{n \geq 1}$, $a = 1, 2$, are generated by a sequence of i.i.d. Bernoulli random variables has found that results on the convergence of X_n depend on the behaviour of the deterministic step size sequences $(\gamma_n)_{n \geq 1}$, generally finding that the quicker they converge to zero, the better behaved the algorithm, although the slower the rate of convergence (recall that we require that $\sum_{n \geq 1} \gamma_n = \infty$ and so γ_n cannot converge to 0 arbitrarily fast). For example, under no assumptions on the learning sequence other than that $\Gamma_n \to \infty$ the algorithm converges to a $\{0, 1\}$-valued random variable but is not guaranteed to be consistent; however under (**S2**) the algorithm is consistent [4].

The results in this chapter are stated under various combinations of assumptions involving just the payoff, or both the step and payoff sequences:

(E1) (Eϕ) is satisfied for some non-decreasing ϕ which is concave on $[k_0, \infty)$ for some $k_0 \in \mathbb{N}$, and $\sup_{n\in\mathbb{N}} \gamma_n \phi(n) < \infty$;

(E2) $(E\phi)$ holds with

$$\phi(n) = n(\log n)^{-(1+\epsilon)}$$

for some $\epsilon > 0$, and $\sup_{n\in\mathbb{N}} \gamma_n \phi(n) < \infty$;

(E3) (**E3.1**) and (**E3.2**) hold;

(E3.1) $\Phi^1_{n,\gamma} := \sum_{k=1}^n (p_1 - \eta_{1,n})\gamma_k$ converges;
(E3.2) $\Phi^2_{n,\gamma} := \sum_{k=1}^n (p_2 - \eta_{2,n})\gamma_k$ converges;

(E4) $(\eta_{2,n} = 1) \implies (\eta_{1,n} = 1)$.

We use a result found in [7] to give conditions under which the sequence $(X_n)_{n\geq 1}$ converges:

Theorem 2.1 *Under assumptions (**S1**) and (**E1**) X_n converges to a $\{0,1\}$-valued random variable. Under these assumptions we write $X_\infty := \lim_{n\to\infty} X_n$.*

Consistency Results We assume throughout the chapter that $p_1 > p_2$ and we say that the Narendra Algorithm is *consistent* if $X_n \to 1$, almost surely, and *inconsistent* otherwise.

In [7] it is proved that under conditions (**S2**) and (**E2**) the algorithm is consistent. Furthermore they prove that the algorithm is inconsistent if $\sum_{n\geq 1}\prod_{k=1}^n (1 - \eta_{2,k}\gamma_k) < \infty$.

In this work we are able to show under different assumptions a more complete characterisation of consistency, generalising that proved in [2]:

Theorem 2.2 *1. Suppose that (**S1**), (**S3**), (**E1**) and one of (**E3**), or (**E4**) hold. Then for $x \in (0,1)$,*

$$\mathbb{P}_x(X_\infty = 0) > 0 \iff \sum_{n\geq 1}\prod_{k=1}^n (1 - \eta_{2,k}\gamma_k) < \infty$$

*2. Suppose that (**S5**) and (**E3.2**) hold. Then*

$$\sum_{n\geq 1}\prod_{k=1}^n (1 - p_2\gamma_k) < \infty \iff \sum_{n\geq 1}\prod_{k=1}^n (1 - \eta_{2,k}\gamma_k) < \infty.$$

The new assumptions in this theorem relating to the of determinism of the sequences $(\eta_{a,n})_{n\geq 1}$ are (**E1**), (**E3**), and (**E4**). To understand the limits (**E1**) and (**E3**) impose on the theorem, consider the family of step sequences

$$\left(\gamma_n = \left(\frac{D}{n+D}\right)^{\alpha}\right)_{n\geq 1}, \text{ where } \alpha \in (0,1], \ D > 0.$$

Assuming that the sequences $(\eta_{a,n})_{n\geq 1}$ are generated by i.i.d. Bernoulli random variables, the Law of the Iterated Logarithm gives an asymptotic lower bound on function ϕ which demands that $\alpha > \frac{1}{2}$ for (**E1**) to hold. A more difficult analysis shows that (**E3**) holds when $\alpha > \frac{1}{2}$ (see Section 2.7 for details).

The assumption (**E4**) has the effect of coupling the deterministic reward sequences in a manner analogous to coupling two randomly generated reward sequences (a technique used often in the papers [2] and [4]).

Rates of Convergence Results on the rate of convergence given in [2] also generalise to the ergodic context. Indeed the proof of Theorem 2.2 itself provides the rate of convergence to 0 when the algorithm is inconsistent. This proof hinges on the result that, on the event $\{X_\infty = 0\}$, $U_{n+1} \leq X_n$ only finitely often with probability 1 (Proposition 2.5), and a fast rate of convergence is observed. Thus we obtain the following theorem:

Theorem 2.3 (Rate of Convergence to 0) *Suppose that (**S1**), (**S3**), (**S5**), (**E1**) and (**E3**) hold. Then*

$$X_n \sim {}^1e^{-}p_2\Gamma_n \text{ almost surely on } \{X_\infty = 0\}.$$

However in contrast to the case of convergence to zero, on the event $\{X_\infty = 1\}$ much more complicated behaviour can be observed. To illuminate this behaviour we study the non-negative martingale

$$Y_n := \frac{1 - X_n}{\theta_n}, \text{ where } \theta_n := \prod_{k=1}^{n} \left(1 - \gamma_k \left(\eta_{1,k} - \eta_{2,k}\right) X_{k-1}\right).$$

[1]Let $(a_n)_{n\geq 0}$ and $(b_n)_{n>0}$ be two non-negative sequences. In this chapter we write $a_n = O(b_n)$ if there exists some constant $C \geq 0$ such that $a_n \leq Cb_n$ for all $n \geq 0$. We write $a_n \sim b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $a_n = o(b_n)$ if $\lim_{n\to\infty}(a_n/b_n) = 0$

Since $(Y_n)_{n\geq 1}$ is a non-negative martingale it converges to a limit Y_∞ and in Proposition 2.11 we find that when $Y_\infty > 0$ a slow rate of convergence is observed, but that when $Y_\infty = 0$ a faster rate of convergence must be observed. In Proposition 2.12 we find conditions under which $Y_\infty > 0$ with positive probability, and conditions under which $Y_\infty = 0$ almost surely.

In order to find the form of the fast rate we guess that, as in the case when $X_n \to 0$, it occurs when $U_{n+1} > X_n$ only finitely often, and in Proposition 2.15 we find conditions under which this guess holds true. This not only provides us with the fast rate itself but also with a characterisation of when both rates can occur with positive probability.

We summarise these results on the rates of convergence in the following Theorem:

Theorem 2.4 (Rates of Convergence to 1) *Assume (**S1**), (**E1**) and (**E3**) hold.*

1. *Then*
$$(1 - X_n) \sim e^{-\pi\Gamma_n} \text{ almost surely on } \{Y_\infty > 0\}.$$

2. *Under (**S3'**)*
$$(1 - X_n) \sim e^{-p_1\Gamma_n} \text{ almost surely on } \{Y_\infty = 0\},$$
$$\mathbb{P}(X_\infty = 1, Y_\infty > 0) > 0,$$
and
$$\mathbb{P}(Y_\infty = 0) > 0 \iff \left(\sum_{k\geq 1}\prod_{k=1}^{n}(1 - p_1\gamma_k) < \infty\right).$$

3. *Under (**S4**) and (**S5**)*
$$(1 - X_n) \sim e^{p_1\Gamma_n} \text{ almost surely on } \{X_\infty = 1\}.$$

The rest of this chapter is structured as follows. After recalling some useful results and notations from previous work we outline the proof of the characterisation of consistency in Section 2.3. In Section 2.4 we give the proofs of the new Lemmas needed in the proof of Theorem 2.2. In the Sections 2.5 and 2.6 we give the results on the rates of convergence to 1. Finally we discuss the assumptions in Section 2.7.

2.2 Preliminaries

Adopting the notation of recent literature we rewrite (2.1) as

$$\begin{aligned} X_0 &= x \in [0,1]\,, \\ X_n &= X_{n-1} + \gamma_n \eta_{1,n} \left(1 - X_{n-1}\right) \mathbf{1}_{\{U_n \leq X_{n-1}\}} \\ &\qquad - \gamma_n \eta_{2,n} X_{n-1} \mathbf{1}_{\{U_n > X_{n-1}\}}, \end{aligned} \tag{2.2}$$

As a stochastic approximation procedure we can rewrite (2.2) in the canonical form:

$$X_n = X_{n-1} + \gamma_n \left(\pi f\left(X_{n-1}\right) + \Delta\Lambda_n\right) + \gamma_n \Delta M_n \tag{2.3}$$

where

$$\begin{aligned} \Delta M_n :=& \eta_{1,n} \left(1 - X_{n-1}\right) \left(\mathbf{1}_{\{U_n \leq X_{n-1}\}} - X_{n-1}\right) \\ &- \eta_{2,n} X_{n-1} \left(\mathbf{1}_{\{U_n > X_{n-1}\}} - \left(1 - X_{n-1}\right)\right), \end{aligned}$$

$$\Delta\Lambda_n := f\left(X_{n-1}\right)\left(\psi_n - \pi\right)$$

and

$$f : x \mapsto x(1-x),\ \pi := p_1 - p_2, \psi_n := \eta_{1,n} - \eta_{2,n}.$$

When the payoff sequences are i.i.d. Bernoulli random variables we can incorporate $\Delta\Lambda_n$ into the martingale increment ΔM_n. However since we assume that the payoff sequences are deterministic this is no longer possible and we have to deal with the resulting deterministic part of the algorithm, $\Delta\Lambda_n$. In order to analyse the growth of this term we often appeal to discrete integration by parts which allows us to describe $\Lambda_n := \sum_{1 \leq k \leq n} \Delta\Lambda_k$ in terms of the ergodic estimates guaranteed by the assumption (Eϕ).

We recall the useful result from [7] (property (7) on page 9) that under the assumptions (**E1**) and (**S1**)

$$\sum_{k=1}^{\infty} \gamma_k X_{k-1}\left(1 - X_{k-1}\right) < \infty. \tag{2.4}$$

Throughout the chapter we shall denote by C a possibly random, but strictly positive constant whose value may change from line to line.

2.3 Characterisation of Consistency

The key to proving the characterisation of consistency lies in showing first that on the event that $\{X_\infty = 0\}$ we must have a "fast rate of convergence", i.e. that $\sum_{n\geq 0} X_n < \infty$ almost surely:

Proposition 2.5 *Suppose that (**S1**), (**S3**), (**E1**) and one of (**E3**) or (**E4**) hold. Then*

$$\{X_\infty = 0\} = \left\{ \sum_{n\geq 0} X_n < \infty \right\} \text{ almost surely.}$$

To prove this result for general, deterministic reward sequences we require three new lemmas. The proofs of these Lemmas form the bulk of our contribution to this argument and are found in the proceeding section.

The first Lemma provides machinery for the analysis of the ergodic component, the Λ_n process. It allows us to patch the arguments from [2], but introduces the assumption (**E1**).

Lemma 2.6 *Suppose that (**E1**) holds. Then on the set $\{X_n \to 0\} \cap \{\sum_n X_n = +\infty\}$, for $a = 1, 2$,*

$$\sum_{k=1}^{n} f\left(X_{k-1}\right)\left(\eta_{a,k} - p_a\right) = A_{a,n} + B_{a,n}$$

where

$$\frac{B_{a,n}}{\sum_{k=0}^{n} X_k} \to 0 \text{ as } n \to \infty,$$

and

$$\frac{A_{a,n}}{\sum_{k=0}^{n} X_k} = o\left(\frac{X_n}{\gamma_n \sum_{k=1}^{n} X_{k-1}}\right).$$

In the second Lemma we show that the martingale $M_n := \sum_{1\leq k\leq n} \Delta M_k$ can be treated exactly as the martingale is treated in [2]:

Lemma 2.7 *On the set $\{X_n \to 0\} \cap \{\sum_n X_n = +\infty\}$ we have*

$$\frac{M_n}{\sum_{k=1}^{n} X_{k-1}} \to 0.$$

In the third Lemma we provide a new proof of a result from [4] which can be directly applied in the Ergodic case under the assumption (**E3**). An adaptation of the proof from [4] provides the same result under the assumption (**E4**):

Lemma 2.8 *Suppose* $\sum_{k\geq 0}\gamma_k^2 < +\infty$*, and that either (**E3**) or (**E4**) hold. Then on* $\{X_n \to 0\}$

$$\limsup_{n\to\infty} \frac{X_n}{\sum_{k\geq n}\gamma_{k+1}^2} < +\infty \text{ almost surely.}$$

Proof of Proposition 2.5. The argument is similar to the one developed by Lamberton and Pagès [2].

By Theorem 2.1, from assumptions (**S1**) and (**E1**) we know that X_∞ exists. So suppose that $X_\infty = 0$ and $\sum_n X_n = +\infty$.

First we prove a lower bound on the rate at which X_n can converge to zero under the assumption that $\sum_n X_n = +\infty$:

$$\liminf_{n\to\infty} \frac{X_n}{\gamma_n \sum_{k=1}^n X_{k-1}} > 0 \text{ a.s..} \tag{2.5}$$

To show this, first note that

$$\begin{aligned}\frac{X_{n+1}}{\gamma_{n+1}} &= \frac{X_n}{\gamma_{n+1}} + \pi\left(1 - X_n\right) X_n + \Delta M_{n+1} + \Delta\Lambda_{n+1} \\ &= \frac{X_n}{\gamma_n} + \left(\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} + \pi\left(1 - X_n\right)\right) X_n + \Delta M_{n+1} + \Delta\Lambda_{n+1}\end{aligned}$$

so that by summing up and letting $\gamma_0 = \gamma_1$

$$\frac{X_n}{\gamma_n} = \frac{x}{\gamma_1} + \sum_{k=1}^n \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} + \pi\left(1 - X_{k-1}\right)\right) X_{k-1} + M_n + \Lambda_n.$$

Now from Lemma 2.7 we know that

$$\lim_{n\to\infty} \frac{M_n}{\sum_{k=1}^n X_{k-1}} = 0 \text{ a.s..}$$

Hence (2.5) follows from Lemma 2.6 and the assumptions (**E1**) and (**S3**).

Let $Z_n = \sum_{k=1}^n X_k$. Note $Z_n \sim \sum_{k=1}^n X_{k-1}$, so that, using (2.5),

$$\exists C > 0 \text{ such that } \forall n \geq 1,\ \gamma_n \leq C\frac{X_n}{Z_n} < +\infty.$$

So

$$\sum_n \gamma_n^2 \leq C^2 \sum_n \frac{X_n^2}{Z_n^2} \leq C^2 \sum_n \frac{X_n}{Z_n^2} < \infty.$$

Hence $\{X_\infty = 0\} \cap \{\sum_n X_n = +\infty\}$ must be null when $\sum_n \gamma_n^2 = \infty$.

So suppose that $\sum_n \gamma_n^2 < \infty$. Then Lemma 2.8 gives us an upper bound on the rate at which X_n can converge to zero:

$$\limsup_{n\to\infty} \frac{X_n}{\sum_{k\geq n} \gamma_{k+1}^2} < +\infty \text{ a.s..} \tag{2.6}$$

We find that these two bounds (2.5) and (2.6) lead to a contradiction.

We know from (2.6) that $X_n \leq C \sum_{k\geq n} \gamma_n^2$ for some $C > 0$, and, by plugging in the estimate $\gamma_{k+1} \leq C\frac{X_{k+1}}{Z_{k+1}}$, we derive

$$\begin{aligned} X_n &\leq C \sum_{k\geq n} \frac{X_k^2}{Z_k^2} \\ &\leq C \left(\sup_{k\geq n} X_k\right) \sum_{k\geq n} \frac{X_k}{Z_k^2} \\ &\leq C\frac{\sup_{k\geq n} X_k}{Z_n}. \end{aligned}$$

Now by assumption $\lim_{n\to\infty} Z_n = +\infty$, so, for n large enough, say $n \geq N$, we have

$$X_n \leq \frac{\sup_{k\geq n} X_k}{2}.$$

By taking n to be the largest integer such that $X_n \geq X_N$ (which exists when $X_n \to 0$ because $X_N > 0$), we reach a contradiction. ■

Now from this Proposition we can prove the first part of Theorem 2.2 with a Borel-Cantelli argument.

Proof of Theorem 2.2 part 1. The argument is again similar to the one developed in [2]. All that we need change is to replace the random rewards of [2] with the deterministic rewards we assume here:

From Proposition 2.5 and the Conditional Borel-Cantelli Lemma (see [8], pg. 124), it follows that $\mathbb{P}_x$-a.s.

$$\begin{aligned}\{X_n \to 0\} &= \left\{\sum_{n\geq 0} X_n < +\infty\right\} \\ &= \left\{\sum_{n\geq 0} \mathbf{1}_{\{U_{n+1}\leq X_n\}} < +\infty\right\} = \bigcup_{n\geq 0}\bigcap_{k\geq n}\{U_{k+1} > X_k\}\,.\end{aligned}$$

The sequence of events $\left(\bigcap_{k\geq n}\{U_{k+1} > X_k\}\right)_{n\geq 1}$ being non-decreasing, we have

$$\mathbb{P}_x\left(X_n \to 0\right) = \lim_{n\to\infty} \mathbb{P}_x\left(\bigcap_{k\geq n}\{U_{k+1} > X_k\}\right).$$

Now denote by $\mathcal{F}_n$ the σ-field generated by the random variables $\{U_n\}_{1\leq k\leq n}$. Then from the definition of the sequence (X_n) we get

$$\begin{aligned}&\mathbb{P}_x\left(\bigcap_{k\geq n}\{U_{k+1} > X_k\} \mid \mathcal{F}_n\right) \\ &\qquad = \mathbb{P}_x\left(\bigcap_{k\geq n}\left\{U_{k+1} > X_n \prod_{l=n+1}^{k}(1-\eta_{2,l}\gamma_l)\right\} \mid \mathcal{F}_n\right) \qquad (2.7)\\ &\qquad = \prod_{k=n}^{\infty}\left(1 - X_n \prod_{l=n+1}^{k}(1-\eta_{2,l}\gamma_l)\right)\end{aligned}$$

and the infinite product is positive iff

$$\sum_{n\geq 1}\prod_{k=1}^{n}(1-\eta_{2,k}\gamma_k) < \infty$$

which proves the result. ■

The proof of the second part of the Theorem follows a different argument.

Proof of Theorem 2.2 part 2. Consider

$$\begin{aligned}&\sum_{n=1}^{N}\prod_{k=1}^{n}\left(1-\eta_{2,k}\gamma_k\right)\\&=\sum_{n=1}^{N}\left[\prod_{k=1}^{n}\left(1-p_2\gamma_k\right)\prod_{k=1}^{n}\frac{\left(1-\eta_{2,k}\gamma_k\right)}{\left(1-p_2\gamma_k\right)}\right].\end{aligned}$$

Note that

$$\frac{\left(1-\eta_{2,k}\gamma_k\right)}{\left(1-p_2\gamma_k\right)}=1-\frac{\left(p_2-\eta_{2,k}\right)\gamma_k}{\left(1-p_2\gamma_k\right)}.$$

Now using discrete integration by parts we find that

$$\begin{aligned}\sum_{k=1}^{n}\frac{\left(p_2-\eta_{2,k}\right)\gamma_k}{\left(1-p_2\gamma_k\right)}&=\sum_{k=1}^{n}\frac{1}{1-p_2\gamma_k}\left(\sum_{i=1}^{k}\left(p_2-\eta_{2,i}\right)\gamma_i-\sum_{i=1}^{k-1}\left(p_2-\eta_{2,i}\right)\gamma_i\right)\\&=\frac{1}{1-p_2\gamma_n}\sum_{k=1}^{n}\left(p_2-\eta_{2,k}\right)\gamma_k\\&\quad-\sum_{k=1}^{n-1}\left(\frac{1}{1-p_2\gamma_{k+1}}-\frac{1}{1-p_2\gamma_k}\right)\sum_{i=1}^{k}\left(p_2-\eta_{2,i}\right)\gamma_i.\end{aligned}$$

But

$$\left(1-p_2\gamma_{k+1}\right)^{-1}-\left(1-p_2\gamma_k\right)^{-1}=p_2\left[\left(\gamma_{k+1}-\gamma_k\right)+O\left(\gamma_{k+1}^2\right)-O\left(\gamma_k^2\right)\right].$$

Moreover the assumption (**S1**) implies that $\left(\gamma_{k+1}-\gamma_k\right)\leq 0$ for all $k\geq 1$, and

$$\sum_{k=1}^{\infty}\left(\gamma_{k+1}-\gamma_k\right)=\gamma_1<\infty$$

So

$$\left((\mathbf{S1}),(\mathbf{S5}),(\mathbf{E3.2})\right)\implies\left(\sum_{k=1}^{n}\frac{\left(p_2-\eta_{2,k}\right)\gamma_k}{\left(1-p_2\gamma_k\right)}<\infty\right).$$

Since

$$\sum_{k=1}^{n}\frac{\left(p_2-\eta_{2,k}\right)\gamma_k}{\left(1-p_2\gamma_k\right)}<\infty\implies 0<\prod_{k=1}^{\infty}\frac{\left(1-\eta_{2,k}\gamma_k\right)}{\left(1-p_2\gamma_k\right)}<\infty$$

the result is proved. ■

2.4 Proof of Lemmas 2.6, 2.7 and 2.8

The aim of the proof of Lemma 2.6 is to analyse the growth of the partial sum $\sum_{k=1}^{N} f(X_{k-1})(\eta_{a,k} - p_a)$. In order to do this it is natural to appeal to discrete integration by parts, since we have some control over the partial sums $\sum_{k=1}^{N} f(X_{k-1})$ and $\sum_{k=1}^{N} (\eta_{a,k} - p_a)$.

Proof of Lemma 2.6. Using summation by parts we obtain

$$\begin{aligned}\sum_{k=1}^{n} f(X_{k-1})(\eta_{a,k} - p_a) &= \sum_{k=1}^{n} f(X_{k-1}) \left(\sum_{i=1}^{k} (\eta_{a,i} - p_a) - \sum_{i=1}^{k-1} (\eta_{a,i} - p_a) \right) \\ &= f(X_{n-1}) \sum_{k=1}^{n} (\eta_{a,k} - p_a) - \sum_{k=1}^{n-1} (f(X_k) - f(X_{k-1})) \sum_{i=1}^{k} (\eta_{a,i} - p_a).\end{aligned}$$

First we investigate the last term in this equation. Using the fact that f is 1-Lipschitz together with (2.2) we find that

$$\begin{aligned}|f(X_k) - f(X_{k-1})| &\leq |X_k - X_{k-1}| \\ &\leq \gamma_k \left(\mathbf{1}_{\{U_k \leq X_{k-1}\}} + X_k \right)\end{aligned}$$

So using the assumptions that (**E1**) holds

$$\left| \sum_{k=1}^{n-1} (f(X_k) - f(X_{k-1})) \sum_{i=1}^{k} (\eta_{a,i} - p_a) \right| \leq \sum_{k=1}^{n-1} \gamma_k \phi(k) r_{a,k} \left(\mathbf{1}_{\{U_k \leq X_{k-1}\}} + X_k \right)$$

where

$$r_{a,n} := \frac{\sum_{k=1}^{n} (\eta_{a,k} - p_a)}{\phi(n)}.$$

Hence from the assumption that $\sum_{k=1}^{\infty} X_k = \infty$ we deduce from the Strong Law of Large Numbers for martingales (see for example [8], page 123)

$$\left| \frac{\sum_{k=1}^{n-1} (f(X_k) - f(X_{k-1})) \sum_{i=1}^{k} (\eta_{a,i} - p_a)}{\sum_{k=1}^{n} X_k} \right| = o(1).$$

The Lemma follows after observing that

$$\frac{X_n}{X_{n-1}} \geq 1 - \gamma_n$$

and that

$$f\left(X_{n-1}\right)\sum_{1}^{n}\left(\eta_{a,k}-p_{a}\right)=\frac{X_{n-1}}{\gamma_{n}}\gamma_{n}\phi\left(n\right)r_{a,n}\left(1-X_{n-1}\right).$$

■

The proof of Lemma 2.7 is little more than a calculation:

Proof of Lemma 2.7. Recall that

$$\begin{aligned}\Delta M_{n}:=&\eta_{1,n}\left(1-X_{n-1}\right)\left(\mathbf{1}_{\{U_{n}\leq X_{n-1}\}}-X_{n-1}\right)\\&-\eta_{2,n}X_{n-1}\left(\mathbf{1}_{\{U_{n}>X_{n-1}\}}-\left(1-X_{n-1}\right)\right)\\=&\left(\eta_{1,n}\left(1-X_{n-1}\right)+\eta_{2,n}X_{n-1}\right)\left(\mathbf{1}_{\{U_{n}\leq X_{n-1}\}}-X_{n-1}\right).\end{aligned}$$

So

$$\begin{aligned}&\mathbb{E}[\left(\Delta M_{n}\right)^{2}\mid\mathcal{F}_{n-1}]\\&\qquad=\left[\eta_{1,n}\left(1-X_{n-1}\right)+\eta_{2,n}X_{n-1}\right]^{2}f\left(X_{n-1}\right).\end{aligned}\tag{2.8}$$

Hence on $\{X_{\infty}=0\}\cap\{\sum_{n}X_{n}=+\infty\}$

$$\left\langle M\right\rangle_{n}=O\left(\sum_{k=1}^{n}X_{k-1}\right).$$

So we can deduce from the Strong Law of Large Numbers for martingales [8] that on $\{X_{\infty}=0\}\cap\{\sum_{n}X_{n}=+\infty\}$

$$\frac{M_{n}}{\sum_{k=1}^{n}X_{k-1}}\rightarrow0$$

as $n\rightarrow\infty$. ■

The proof that we use for Lemma 2.8 depends on the assumptions that we make. We give one proof under the assumption (**E3**), and a different proof under the assumption (**E4**):

Proof of Lemma 2.8 Under (E3). Note that

$$\begin{aligned}X_{n}&=X_{n-1}+\gamma_{n}\pi f\left(X_{n-1}\right)+\gamma_{n}\Delta\Lambda_{n}+\gamma_{n}\Delta M_{n}\\&=X_{n-1}\left(1+\gamma_{n}\psi_{n}\left(1-X_{n-1}\right)\right)+\gamma_{n}\Delta M_{n}\end{aligned}$$

So let $\bar{\theta}_n^{(n)} := 1$ and, for $k > n$, let

$$\bar{\theta}_k^{(n)} := \prod_{i=n+1}^{k} \left(1 + \gamma_i \psi_i \left(1 - X_{i-1}\right)\right).$$

Define, for $k \geq n$,

$$Z_k^{(n)} := \frac{X_k}{\bar{\theta}_k^{(n)}}.$$

Then $\left(Z_k^{(n)}\right)_{k \geq n}$ is a martingale.

Define also, for $1 > \mu > 0$, the stopping time

$$T_\mu^{(n)} := \inf\left\{k \geq n : \bar{\theta}_k^{(n)} < \mu\right\}.$$

Now for any sequence of numbers $(a_n)_{n \geq 1} \in (-1, 1)^{\mathbb{N}}$ such that $\sum_{n \geq 1} a_n^2 < \infty$

$$\prod_{n \geq 1} (1 + a_n) = +\infty \iff \sum_{n \geq 1} a_n = +\infty.$$

So, since we have assumed that $\sum_{n \geq 1} \gamma_n^2 < \infty$, (**E3**) implies that on $\{X_\infty = 0\}$

$$\bar{\theta}_k^{(n)} = \prod_{i=n+1}^{k} \left(1 + \gamma_i \psi_i \left(1 - X_{i-1}\right)\right) \to +\infty \text{ a.s.}$$

as $k \to \infty$. Hence on $\{X_\infty = 0\} \cap \left\{T_\mu^{(n)} = \infty\right\}$,

$$Z_{\infty \wedge T_\mu^{(n)}}^{(n)} := \lim_{k \to \infty} Z_{k \wedge T_\mu^{(n)}}^{(n)} = 0$$

almost surely.

Now

$$\begin{aligned}
\mathbb{P}_x\left(\{X_\infty = 0\} \cap \left\{T_\mu^{(n)} = \infty\right\} \mid \mathcal{F}_n\right) &\leq \mathbb{P}_x\left(\left\{Z^{(n)}_{\infty \wedge T_\mu^{(n)}} = 0\right\} \mid \mathcal{F}_n\right) \\
&= \mathbb{E}_x\left(\mathbf{1}_{\left\{Z^{(n)}_{\infty \wedge T_\mu^{(n)}} = 0\right\}} \mid \mathcal{F}_n\right) \\
&= \frac{\mathbb{E}_x\left(\mathbf{1}_{\left\{Z^{(n)}_{\infty \wedge T_\mu^{(n)}} = 0\right\}} \left(Z^{(n)}_{n \wedge T_\mu^{(n)}}\right)^2 \mid \mathcal{F}_n\right)}{\left(Z^{(n)}_{n \wedge T_\mu^{(n)}}\right)^2} \\
&\leq \frac{\mathbb{E}_x\left(\left(Z^{(n)}_{\infty \wedge T_\mu^{(n)}} - Z^{(n)}_{n \wedge T_\mu^{(n)}}\right)^2 \mid \mathcal{F}_n\right)}{X_n^2}.
\end{aligned}$$

But

$$\begin{aligned}
\mathbb{E}_x\left(\left(Z^{(n)}_{\infty \wedge T_\mu^{(n)}} - Z^{(n)}_{n \wedge T_\mu^{(n)}}\right)^2 \mid \mathcal{F}_n\right) &= \mathbb{E}_x\left(\sum_{k=n+1}^{T_\mu^{(n)}} \left(\frac{\gamma_k \Delta M_k}{\bar{\theta}_k^{(n)}}\right)^2 \mid \mathcal{F}_n\right) \\
&\leq \frac{1}{\mu}\mathbb{E}_x\left(\sum_{k=n+1}^{T_\mu^{(n)}} \frac{\left(\gamma_k \Delta M_k\right)^2}{\bar{\theta}_k^{(n)}} \mid \mathcal{F}_n\right) \\
&\leq \frac{1}{\mu}\mathbb{E}_x\left(\sum_{k=n+1}^{\infty} \frac{\left(\gamma_k \Delta M_k\right)^2}{\bar{\theta}_k^{(n)}} \mid \mathcal{F}_n\right) \\
&= \frac{1}{\mu}\mathbb{E}_x\left(\sum_{k=n+1}^{\infty} \frac{\gamma_k^2}{\bar{\theta}_k^{(n)}} \mathbb{E}_x\left(\left(\Delta M_k\right)^2 \mid \mathcal{F}_{k-1}\right) \mid \mathcal{F}_n\right) \\
&\leq \frac{C}{\mu}\mathbb{E}_x\left(\sum_{k=n+1}^{\infty} \gamma_k^2 Z_{k-1}^{(n)} \mid \mathcal{F}_n\right) \\
&= \frac{C}{\mu}\sum_{k=n+1}^{\infty} \gamma_k^2 \mathbb{E}_x\left(Z_{k-1}^{(n)} \mid \mathcal{F}_n\right) \\
&= \frac{C}{\mu} X_n \left[\sum_{k=n+1}^{\infty} \gamma_k^2\right].
\end{aligned}$$

So

$$\mathbb{P}_x\left(\{X_\infty = 0\} \cap \left\{T_\mu^{(n)} = \infty\right\} \mid \mathcal{F}_n\right) \leq \frac{\frac{C}{\mu} X_n \sum_{k=n+1}^{\infty} \gamma_k^2}{X_n^2}$$
$$= \frac{C}{\mu} \frac{\sum_{k=n+1}^{\infty} \gamma_k^2}{X_n}$$

The result now follows from the fact that

$$\mathbb{P}_x\left(\{X_\infty = 0\} \cap \left\{T_\mu^{(n)} = \infty\right\} \mid \mathcal{F}_n\right) \to \mathbf{1}_{\{X_\infty = 0\}}$$

as $n \to \infty$. ■

Remark 2.9 *The proof of Lemma 2.8 under assumption (**E3**) varies from the proof in [4] in that it uses a different martingale to make the analysis. We use here the analogue of the martingale used to analyse the case of convergence to* 1*, which we denote by* Y_n *in the proceeding section.*

Using this martingale it is possible to generalise the results obtained in [2] for the consistency and rate of convergence to 0 *in the Bernoulli Case to a Bandit with continuous, bounded i.i.d. payoffs, removing the dependence of the argument on a coupling of the reward sequences. The requirement that* $\bar{\theta}_k^{(n)} \to \infty$ *becomes a trivial consequence of the assumption that* $\sum_{k \geq 1} \gamma_k = \infty$ *(i.e. the assumption (**E3**) is not needed).*

Proof of Lemma 2.8 Under (E4). Suppose that the reward sequences $(\eta_{a,n})_{n \geq 1}$, $a = 1, 2$, satisfy $(\eta_{2,n} = 1 \implies \eta_{1,n} = 1)$. Then we may define the bounded martingale $\left(X_k^{(n)}\right)_{k \geq n}$ by $X_n^{(n)} := X_n$ and, for $k > n$,

$$X_k^{(n)} := X_{k-1}^{(n)} + \gamma_k \eta_{2,k} \left(\mathbf{1}_{\{U_{k-1} \leq X_{k-1}\}} - X_{k-1}\right).$$

Now by inspection we find that $X_k^{(n)} < X_k$ for all $k \geq n$, and so $X_\infty = 0 \implies X_\infty^{(n)} = 0$.

So, as before, we can deduce that

$$\begin{aligned}\mathbb{P}_x\left(X_\infty = 0 \mid \mathcal{F}_n\right) &\leq \mathbb{P}_x\left(X_\infty^{(n)} = 0 \mid \mathcal{F}_n\right) \\ &\leq \frac{\mathbb{E}_x\left(\left(X_\infty^{(n)} - X_n^{(n)}\right)^2 \mid \mathcal{F}_n\right)}{X_n^2} \leq \frac{\mathbb{E}_x\left(\sum_{k=n+1}^\infty \Delta\left\langle X^{(n)}\right\rangle_k \mid \mathcal{F}_n\right)}{X_n^2} \\ &\leq \frac{\sum_{k=n+1}^\infty \gamma_k^2 \mathbb{E}_x\left(f\left(X_k\right) \mid \mathcal{F}_n\right)}{X_n^2} \leq \frac{\sum_{k=n+1}^\infty \gamma_k^2 \mathbb{E}_x\left(X_k \mid \mathcal{F}_n\right)}{X_n^2} \\ &= \frac{\sum_{k=n+1}^\infty \gamma_k^2}{X_n}\end{aligned}$$

We conclude once again from the fact that $\mathbb{P}_x\left(X_\infty = 0 \mid \mathcal{F}_n\right) \to \mathbf{1}_{\{X_\infty = 0\}}$ as $n \to \infty$. ■

2.5 Rate of Convergence to 0

We can now draw out of the proof of Theorem 2.2 the rate of convergence when the algorithm fails.

Proposition 2.10 *1. Suppose that (**S1**), (**S3**), (**E1**) and one of (**E3**) or (**E4**) hold. On the event $\{X_\infty = 0\}$ there exists a random, but almost surely finite time n_0 such that $\forall\ n \geq n_0$*

$$X_n = X_{n_0} \prod_{k=n_0}^{n} \left(1 - \eta_{2,k}\gamma_k\right);$$

*2. Under the additional assumptions (**S5**) and (**E3.2**) there is a random variable $\Xi_x > 0$ such that*

$$X_n \sim \Xi_x \prod_{k=1}^{n} \left(1 - p_2\gamma_k\right) \ \text{a.s. on } \left\{X_\infty = 0\right\}.$$

Proof. It follows from Proposition 2.5 that under the assumptions (**S1**), (**S3**), (**E1**) and one of (**E3**) or (**E4**), on the event $X_\infty = 0$, after some random, but almost surely finite time, n_0, the strategy plays only implements the second action. Hence part 1. of the Proposition now follows from the definition of the algorithm.

Part 2 now follows easily from Part 1. and the proof of the second part of Theorem 2.2 in which we show that if $\sum \gamma_n^2 < \infty$ and $\Phi_{n,\gamma}^2$ converges, then

$$\frac{\prod_{k=1}^{n}(1-\eta_{2,k}\gamma_k)}{\prod_{k=1}^{n}(1-p_2\gamma_k)} \to \xi < \infty$$

for some non-zero random variable ξ, almost surely. ■

2.6 Rate of Convergence to 1

We quickly set some notation: Let

$$\Gamma'_n = \sum_{k=1}^{n} \psi_k \gamma_k.$$

Note that we can rewrite the algorithm as

$$(1-X_{n+1}) = (1-X_n)(1-\gamma_{n+1}\psi_{n+1}X_n) - \gamma_{n+1}\Delta M_{n+1}.$$

So in analogy with [2] we define

$$\theta_n := \prod_{k=1}^{n}(1-\gamma_k\psi_k X_{k-1}) \text{ and } Y_n := \frac{(1-X_n)}{\theta_n}$$

Then since θ_n is previsible, Y_n is a non-negative martingale, and so it has an almost sure limit Y_∞.

Proposition 2.11 *Suppose that (**E1**) and (**S1**) hold. Then on the set* $\{X_\infty = 1\}$

$$\lim_{n\to\infty} \frac{\theta_n}{\prod_{k=1}^{n}(1-\psi_k\gamma_k)} = \xi$$

almost surely, where ξ *is a finite, positive random variable.*

Proof. Note that

$$\left| \sum_{1\le k\le n} \frac{\gamma_k\psi_k(1-X_{k-1})}{1-\psi_k\gamma_k} \right| \le \sum_{1\le k\le n} \frac{\gamma_k(1-X_{k-1})}{1-\gamma_k}.$$

However (2.4) holds under the assumptions (**E1**) and (**S1**) and from this we deduce that on $\{X_\infty = 1\}$,

$$\sum_{k=1}^{\infty} \left| \frac{\gamma_k \psi_k \left(1 - X_{k-1}\right)}{1 - \psi_k \gamma_k} \right| < \infty$$

almost surely, and so

$$\prod_{k \geq 1} \frac{1 - \gamma_k \psi_k X_{k-1}}{1 - \gamma_k \psi_k} = \prod_{k \geq 1} \left[\frac{(1 - \gamma_k \psi_k) + \gamma_k \psi_k \left(1 - X_{k-1}\right)}{1 - \gamma_k \psi_k} \right] > 0$$

almost surely. The result follows immediately. ■

Proposition 2.11, as in the case of i.i.d. Bernoulli payoffs, implies that when $\sum \gamma_k^2 < \infty$, $1 - X_n = O\left(e^{-\Gamma'_n}\right)$. However we also see that the rates of convergence are directly affected, as seen in [2], by whether $Y_\infty = 0$ or not, and the following Proposition describes conditions under which we can make statements about Y_∞

Proposition 2.12 *Assume that (**S1**), (**E1**) and (**E3**) hold.*

1. *If* $\sum_n \gamma_n^2 e^{\pi \Gamma_n} < \infty$

$$\mathbb{P}\left(X_\infty = 1, Y_\infty > 0\right) > 0.$$

Moreover, on the set $\{Y_\infty = 0\}$,

$$\limsup_{n \to \infty} \frac{Y_n}{\sum_{k \geq n} \gamma_k^2 e^{\Gamma'_k}} < \infty$$

almost surely.

2. *If*

$$\sum_n \gamma_n^2 e^{\pi \Gamma_n} = +\infty, \sum_n \gamma_n^2 < \infty \text{ and } \sup_{n \geq 1} \gamma_n e^{\pi \Gamma_n} < +\infty$$

then, for every $x \in (0, 1)$,

$$\{X_\infty = 1\} = \{Y_\infty = 0\}$$

almost surely

Proof. Assume that (**S1**), (**E1**) and (**E3**) hold, and note that (**E3**) implies

$$\left(\sum_n \gamma_n^2 e^{\pi\Gamma_n} = +\infty\right) \iff \left(\sum_n \gamma_n^2 e^{\Gamma'_n} = +\infty\right),$$

$$\left(\sum_n \gamma_n^2 e^{\pi\Gamma_n} < +\infty\right) \iff \left(\sum_n \gamma_n^2 e^{\Gamma'_n} < +\infty\right),$$

and

$$\left(\sup_{n\geq 1} \gamma_n e^{\pi\Gamma_n} < +\infty\right) \iff \left(\sup_{n\geq 1} \gamma_n e^{\Gamma'_n} < +\infty\right).$$

Notice that from Proposition 2.11 it follows that, when $\sum_n \gamma_n^2 < \infty$,

$$\theta_n e^{\Gamma'_n} \text{ converges as } n \to \infty. \tag{2.9}$$

1. Assume that $\sum_n \gamma_n^2 e^{\pi\Gamma_n} < \infty$.

   First we show that Y_n is bounded in L^2 by estimating its conditional variance process. Note that

   $$\mathbb{E}\left((Y_{n+1} - Y_n)^2 \mid \mathcal{F}_n\right) = \left(\frac{\gamma_n}{\theta_n}\right)^2 \mathbb{E}\left((\Delta M_n)^2 \mid \mathcal{F}_n\right)$$

   and recall from the proof of Lemma 2.7 that

   $$\begin{aligned}\mathbb{E}\left((\Delta M_n)^2 \mid \mathcal{F}_{n-1}\right) &= \left[\eta_{1,n}\left(1 - X_{n-1}\right) + \eta_{2,n} X_{n-1}\right]^2 f\left(X_{n-1}\right) \\ &\leq f\left(X_{n-1}\right).\end{aligned}$$

   Now $\sum_n \gamma_n^2 e^{\pi\Gamma_n} < \infty \implies \sum_n \gamma_n^2 < \infty$ it follows from (2.9) that

   $$\mathbb{E}\left((Y_{n+1} - Y_n)^2 \mid \mathcal{F}_n\right) \leq C\gamma_{n+1}^2 e^{\Gamma'_{n+1}} Y_n. \tag{2.10}$$

   Note that, since $(Y_n)_{n\geq 1}$ is a non-negative martingale, $\sup_n \left(\mathbb{E}\left(Y_n\right)\right) < \infty$ almost surely. So

   $$\left(\sum_n \gamma_n^2 e^{\Gamma'_n} < \infty\right) \implies \left((Y_n)_{n\geq 1} \text{ is bounded in } L^2\right).$$

Next consider the conditional covariance:

$$\begin{aligned}
&\mathbb{E}(f(X_{n-1}) \mid \mathcal{F}_{n-1}) \\
&\quad = (X_{n-1} + \gamma_n \psi_n f(X_{n-1})) \\
&\qquad\qquad . ((1 - X_{n-1})(1 - \gamma_n \psi_n X_{n-1})) - \mathbb{E}\left((\Delta M_n)^2 \mid \mathcal{F}_{n-1}\right) \\
&\quad = f(X_{n-1}) \left[(1 + \gamma_n \psi_n (1 - X_{n-1})) . (1 - \gamma_n \psi_n X_{n-1})\right. \\
&\qquad\qquad \left. -\gamma_n^2 \left[\eta_{1,n}(1 - X_{n-1}) + \eta_{2,n} X_{n-1}\right]^2\right] \\
&\quad = f(X_{n-1}) \left[1 + \gamma_n \psi_n ((1 - X_{n-1}) - X_{n-1})\right. \\
&\qquad\qquad \left. -\gamma_n^2 \left(f(X_{n-1}) \psi_n^2 + \left[\eta_{1,n}(1 - X_{n-1}) + \eta_{2,n} X_{n-1}\right]^2\right)\right] \\
&\quad \geq f(X_{n-1}) \left[1 - \gamma_n \psi_n X_{n-1}\right. \\
&\qquad\qquad \left. +\gamma_n \psi_n (1 - X_n) - 2\gamma_n^2\right] \\
&\quad = f(X_{n-1})(1 - \gamma_n \psi_n X_{n-1}) \left(1 - \frac{2\gamma_n^2 - \gamma_n \psi_n (1 - X_n)}{(1 - \gamma_n \psi_n X_{n-1})}\right) \\
&\quad \geq f(X_{n-1})(1 - \gamma_n \psi_n X_{n-1}) \left(1 - \frac{2\gamma_n^2 + \gamma_n (1 - X_n)}{(1 - \gamma_n)}\right)
\end{aligned}$$

So

$$\mathbb{E}(X_n Y_n \mid \mathcal{F}_{n-1}) \geq X_{n-1} Y_{n-1} \left(1 - \frac{2\gamma_n^2 + \gamma_n (1 - X_n)}{(1 - \gamma_n)}\right)$$

Now for sufficiently large n, say $n \geq n_0$,

$$\left(1 - \frac{2\gamma_n^2 + \gamma_n (1 - X_n)}{(1 - \gamma_n)}\right) \leq 1$$

and so by induction

$$\mathbb{E}(X_n Y_n) \geq \mathbb{E}(X_{n_0} Y_{n_0}) \prod_{k=n_0}^{n} \left(1 - \frac{2\gamma_n^2 + \gamma_n (1 - X_n)}{(1 - \gamma_n)}\right)$$

Now it can be easily deduced from display (6) on page 9 of [7] that, under assumptions (**E1**), (**S1**), and $\sum_{n \geq 0} \gamma_n^2 < \infty$, $(X_n)_{n \geq 0}$ converges in L^2. Moreover we know that $Y_n \to Y_\infty$ in L^2, and since $\sum_n \gamma_n^2 e^{\Gamma'_n} < \infty \implies \sum_n \gamma_n^2 < \infty$ we can conclude, using (2.4), that $\mathbb{E}(X_\infty Y_\infty) > 0$. Hence $\mathbb{P}(X_\infty = 1, Y_\infty > 0) > 0$.

Now we consider the special case when $Y_\infty = 0$. Note that

$$\begin{aligned}\mathbb{P}\left(Y_\infty = 0 \mid \mathcal{F}_n\right) &= \frac{1}{Y_n^2}\mathbb{E}\left(\mathbf{1}_{\{Y_\infty=0\}}\left(Y_\infty - Y_n\right)^2 \mid \mathcal{F}_n\right) \\ &\leq \frac{1}{Y_n^2}\sum_{k\geq n+1}\mathbb{E}\left(\left(Y_{k+1} - Y_k\right)^2 \mid \mathcal{F}_n\right).\end{aligned}$$

If we plug in our estimate (2.10) then we obtain

$$\mathbb{P}\left(Y_\infty = 0 \mid \mathcal{F}_n\right) \leq \frac{C}{Y_n}\sum_{k\geq n+1}\gamma_k^2 e^{\Gamma'_k}$$

where we have used that $\mathbb{E}\left(Y_k \mid \mathcal{F}_n\right) = Y_n$ whenever $k > n$. However $\mathbb{P}\left(Y_\infty = 0 \mid \mathcal{F}_n\right)$ converges almost surely to $\mathbf{1}_{\{Y_\infty=0\}}$, and so the result in part 1 follows.

2. Assume that

$$\sum_n \gamma_n^2 e^{\pi\Gamma_n} = +\infty, \sum_n \gamma_n^2 < \infty \text{ and } \sup_{n\geq 1}\gamma_n e^{\pi\Gamma_n} < +\infty$$

Since once again $\sum_{n\geq 0}\gamma_n^2 < \infty$ we deduce from (2.9) that the martingale $(Y_n)_{n\geq 0}$ has bounded increments:

$$\begin{aligned}\left|Y_n - Y_{n-1}\right| &= \frac{\gamma_n}{\theta_n}\left|\Delta M_n\right| \\ &\leq C\gamma_n e^{\pi\Gamma_n}\left|\Delta M_n\right|\end{aligned}$$

Hence the Law of the Iterated Logarithm for martingales (see for example [1]) implies that $\liminf_n Y_n = -\infty$ on the event $\{\langle Y\rangle_\infty = +\infty\}$. However Y_n is non-negative and so we deduce that $\{\langle Y\rangle_\infty < +\infty\}$ almost surely.

Recall that

$$\begin{aligned}\mathbb{E}\left(\left(\Delta M_k\right)^2 \mid \mathcal{F}_{k-1}\right) &= \left[\eta_{1,k}\left(1 - X_{k-1}\right) + \eta_{2,k}X_{k-1}\right]^2 f\left(X_{k-1}\right) \\ &\geq \left[\eta_{1,k}\left(1 - X_{k-1}\right)^2 + \eta_{2,k}X_{k-1}^2\right] f\left(X_{k-1}\right)\end{aligned}$$

Hence on $\{X_\infty = 1\}$

$$\begin{aligned}\Delta \langle Y \rangle_n &= \frac{\gamma_n^2}{\theta_n^2}\mathbb{E}\left((\Delta M_n)^2 \mid \mathcal{F}_{n-1}\right) \\ &\geq \frac{\gamma_n^2}{\theta_n\left(1-\psi_n\gamma_n X_{n-1}\right)}\left[\eta_{1,n}\left(1-X_{n-1}\right)^2 + \eta_{2,n}X_{n-1}^2\right] X_{n-1}Y_{n-1} \\ &\geq C\left[\eta_{1,n}\left(1-X_{n-1}\right)^2 + \eta_{2,n}X_{n-1}^2\right] X_{n-1}Y_{n-1}\gamma_n^2 e^{\pi\Gamma_n}.\end{aligned}$$

So summing over $n > N$,

$$\begin{aligned}\langle Y \rangle_\infty - \langle Y \rangle_N \geq C\sum_{n\geq N}\eta_{2,n}X_{n-1}^3 Y_{n-1}\gamma_n^2 e^{\pi\Gamma_n} \\ + C\sum_{n\geq N} X_{n-1}Y_{n-1}\gamma_n^2 e^{\pi\Gamma_n}\eta_{1,n}\left(1-X_{n-1}\right)^2\end{aligned} \tag{2.11}$$

Moreover recall that under (**E1**) from (2.4), $\sum \gamma_n\left(1-X_n\right)$ almost surely. It follows that since $\sup_{n\geq 1}\gamma_n e^{\Gamma_n'} < +\infty$ and Y_n converges almost surely, on the event $\{X_\infty = 1\}$

$$C\sum_{n\geq N} X_{n-1}Y_{n-1}\gamma_n^2 e^{\pi\Gamma_n}\eta_{1,n}\left(1-X_{n-1}\right)^2 < \infty.$$

Now, when $X_\infty = 1$ and $Y_\infty > 0$, for sufficiently large N

$$\sum_{n\geq N}\eta_{2,n}X_{n-1}^3 Y_{n-1}\gamma_n^2 e^{\pi\Gamma_n} \geq C\left(\sum_{n\geq N} p_2\gamma_n^2 e^{\pi\Gamma_n} + \sum_{n\geq N}\left(\eta_{2,n}-p_2\right)\gamma_n\left(\gamma_n e^{\pi\Gamma_n}\right)\right).$$

However it follows from (**E3.2**), the assumption that $\sup_{n\geq 1}\gamma_n e^{\pi\Gamma_n} < +\infty$, and Lemma 2.17 (see section 2.7) that

$$\sum_{n\geq N}\left(\eta_{2,n}-p_2\right)\gamma_n^2 e^{\pi\Gamma_n} < \infty.$$

Hence if $\sum \gamma_n^2 e^{\pi\Gamma_n} = +\infty$ we deduce from (2.11) that $Y_\infty = 0$ on the event $\{X_\infty = 1\}$.

■

For the case when $X_n \to 0$ we found that $\sum_{n\geq 1} X_n < \infty$ almost surely and that a fast rate of convergence could be deduced as a result. For the case when $X_n \to 1$ we can characterise the event that $\sum_{n\geq 1}\left(1-X_n\right) < \infty$ and deduce a fast rate of convergence on this event:

Proposition 2.13 *For all* $x \in (0,1)$

$$\mathbb{P}_x\left(\sum_n (1-X_n) < \infty\right) > 0 \iff \sum_{n\geq 1}\prod_{k=1}^{n}(1-\eta_{1,k}\gamma_k) < \infty.$$

Suppose also that under (***S5***) *and* (***E3.1***)

$$\sum_{n\geq 1}\prod_{k=1}^{n}(1-p_1\gamma_k) < \infty \iff \sum_{n\geq 1}\prod_{k=1}^{n}(1-\eta_{1,k}\gamma_k) < \infty.$$

Proof. The arguments employed in the proof of Theorem 2.2 can be applied here also. ■

Corollary 2.14 *On the event* $\{X_\infty = 1\} \cap \left\{\sum_{n\geq 0}(1-X_n) < \infty\right\}$ *there exists an* n_0 *such that* $\forall n \geq n_0$

$$1 - X_n = (1 - X_{n_0}) \prod_{k=n_0}^{n} (1 - \eta_{1,k}\gamma_k).$$

Moreover, under the assumptions (***S5***) *and* (***E3.1***), $\forall n \geq n_0$

$$1 - X_n = (1 - X_{n_0}) \prod_{k=n_0}^{n} (1 - p_1\gamma_k)$$

Proof. The arguments employed in the proof of Proposition 2.10 can be applied here also. ■

Finally in the proposition below we give conditions under which the above results give a complete description of the rates of convergence on the set $\{X_\infty = 1\}$. This proposition is weaker than the analogue found in [2] (Lamberton and Pages do not require (**S1**) to hold, nor (**S5**) to hold in part 1), and the arguments used here are similar.

Proposition 2.15 *Assume* (***S1***), (***E1***) *and* (***E3***) *hold.*

1. *Under* (***S4***) *and* (***S5***)

$$\sum_n (1-X_n) < \infty \text{ almost surely on } \{X_\infty = 1\}.$$

*2. Under (**S3'**)*

$$\sum_n \gamma_n^2 e^{\pi\Gamma_n} < \infty$$

and

$$\sum_n (1 - X_n) < \infty \text{ almost surely on } \{Y_\infty = 0\}.$$

Let us denote

$$\epsilon_n := \frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - \pi.$$

We prove a Lemma which we shall use in the proof of the second part.

Lemma 2.16 *Assume (**E1**) and (**S3'**) hold. Then*

$$\left(Z_n = \frac{(1 - X_n)}{\gamma_n}\right)_{n \geq n_0}$$

is a submartingale, and

$$\sum_n (1 - X_n) < \infty \text{ almost surely on } \{X_\infty = 1\} \cap \left\{\sup_n \frac{1 - X_n}{\gamma_n} < \infty\right\}$$

Proof. Note that

$$\begin{aligned}
\frac{1 - X_n}{\gamma_n} &= \frac{1 - X_{n-1}}{\gamma_n} - \pi f(X_{n-1}) - \Delta\Lambda_n - \Delta M_n \\
&= \frac{1 - X_{n-1}}{\gamma_{n-1}} \left(\frac{\gamma_{n-1}}{\gamma_n} - \gamma_{n-1}\pi X_{n-1}\right) - \Delta\Lambda_n - \Delta M_n \\
&= \frac{1 - X_{n-1}}{\gamma_{n-1}} \left(1 + \gamma_{n-1}\left(\frac{1}{\gamma_n} - \frac{1}{\gamma_{n-1}}\right) - \gamma_{n-1}\pi X_{n-1}\right) \\
&\qquad - \Delta\Lambda_n - \Delta M_n
\end{aligned}$$

So

$$\begin{aligned}
\frac{1 - X_n}{\gamma_n} = \frac{1 - X_0}{\gamma_0} + \sum_{k=1}^{n} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} - \pi X_{k-1}\right)(1 - X_{k-1}) \\
- \Lambda_n - M_n.
\end{aligned}$$

Now suppose $X_\infty = 1$, and $\sum_n (1 - X_n) = \infty$, then

$$\liminf_{n\to\infty} \frac{\sum_{k=1}^n \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} - \pi X_{k-1}\right)(1 - X_{k-1})}{\sum_{k=0}^n (1 - X_k)} \geq \liminf_{n\to\infty} \epsilon_n$$

A minor modification to Lemma 2.7 shows that

$$\frac{M_n}{\sum_{k=0}^n (1 - X_k)} \to 0.$$

Furthermore, following a similar argument to that for Lemma 2.6 we find that on the event $\{X_\infty = 1\} \cap \{\sum_n (1 - X_n) = \infty\}$,

$$\lim_{n\to\infty} \left[\frac{\Lambda_n}{\sum_{k=0}^n (1 - X_k)} - \gamma_n \phi(n) X_{n-1} (r_{1,n} - r_{2,n}) \frac{1 - X_{n-1}}{\gamma_n \sum_{k=0}^n (1 - X_{k-1})}\right] = 0$$

where $r_{1,n}$ and $r_{2,n}$ are as defined in the proof of Lemma 2.6. Hence almost surely on $\{X_\infty = 1\}$

$$\left(\sum_{n\geq 0} (1 - X_n) = \infty\right) \implies \left(\liminf_{n\to\infty} \frac{1 - X_n}{\gamma_n \sum_{k=0}^n (1 - X_k)} > 0\right).$$

The result now follows immediately. ■

Now we prove the proposition:

Proof of Proposition 2.15. To see the first part first we note that

$$\begin{aligned}
\frac{1 - X_{n+1}}{\gamma_{n+1}} &= \frac{1 - X_n}{\gamma_{n+1}} - \pi f(X_n) - \Delta\Lambda_{n+1} - \Delta M_{n+1} \\
&= \frac{1 - X_n}{\gamma_n} \left(1 + \gamma_n \left(\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - \pi X_n\right)\right. \\
&\qquad \left. - \gamma_n (\psi_n - \pi) X_n\right) - \Delta M_{n+1} \\
&= \frac{1 - X_n}{\gamma_n} (1 + \gamma_n \epsilon_n + \gamma_n \pi (1 - X_n) \\
&\qquad - \gamma_n (\psi_n - \pi) X_n) - \Delta M_{n+1} \\
&\leq \frac{1 - X_n}{\gamma_n} \left(1 + \gamma_n \epsilon_n^+ + \pi \gamma_n (1 - X_n)\right. \\
&\qquad \left. - \gamma_n (\psi_n - \pi) X_n\right) - \Delta M_{n+1}. \qquad (2.12)
\end{aligned}$$

So

$$\mathbb{E}\left(\frac{1-X_{n+1}}{\gamma_{n+1}} \mid \mathcal{F}_n\right)$$
$$\leq \frac{1-X_n}{\gamma_n}\left(1+\gamma_n\epsilon_n^+ + \gamma_n\pi\left(1-X_n\right) - \gamma_n\left(\psi_n-\pi\right)X_n\right).$$

and so

$$\frac{1-X_n}{\gamma_n\prod_{k=1}^{n-1}\left(1+\gamma_k\epsilon_k^+ + \pi\gamma_k\left(1-X_k\right) - \gamma_k\left(\psi_k-\pi\right)X_k\right)}$$

is a positive supermartingale. Now on $\{X_\infty = 1\}$ we know from (2.4) that

$$\sum \gamma_k\left(1-X_k\right) < \infty,$$

and by assumption

$$\sum_{k\geq 1}\gamma_k^2, \sum_{n\geq 1}\gamma_n\epsilon_n^+, \Phi^1_{\gamma,\infty}, \Phi^2_{\gamma,\infty} < \infty,$$

Hence the convergence of this supermartingale implies that

$$Z_n = \frac{1-X_n}{\gamma_n}$$

converges almost surely.

However from (2.12) we see that

$$\Delta M_{n+1} \leq Z_n - Z_{n+1} + Z_n\left(\gamma_n\epsilon_n + \gamma_n\pi\left(1-X_n\right)\right.$$
$$\left.-\gamma_n\left(\psi_n-\pi\right)X_n\right)$$

and so on $\{X_\infty = 1\}$ $(M_n)_{n\geq 0}$ is bounded with bounded increments. Hence $\langle M\rangle_\infty < \infty$. But now, as in Lemma 2.7, we can compute $\langle M\rangle_n$ and deduce that on $\{X_\infty = 1\}$

$$\langle M\rangle_\infty < \infty \iff \left(\sum_{n=1}^{\infty}\left(1-X_n\right) < \infty\right).$$

Now we prove the second part. The proof follows the scheme of Lamberton and Pagès [2]. Now by (**E3'**) for large enough n

$$\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - \pi \geq \epsilon.$$

So the sequence (γ_n) is non-increasing and for large enough n

$$\gamma_n - \gamma_{n+1} \geq (\pi + \epsilon)\, \gamma_n \gamma_{n+1}$$

and so $\sum_n \gamma_n^2 < \infty$. Also, for large enough n

$$\gamma_{n+1} \leq \gamma_n \left(1 - (\epsilon + \pi)\, \gamma_{n+1}\right) \leq \gamma_n \exp\left(-(\epsilon + \pi)\, \gamma_{n+1}\right)$$

So for $k \geq n$

$$\gamma_k \leq \gamma_n \exp\left(-\left((\pi + \epsilon)\left(\Gamma_k - \Gamma_n\right)\right)\right)$$

and

$$\begin{aligned}
\sum_{k \geq n} \gamma_k^2 e^{\Gamma_k} &\leq \sum_{k \geq n} \gamma_k \gamma_n \exp\left(-(\pi + \epsilon)\left(\Gamma_k - \Gamma_n\right) + \pi \Gamma_k\right) \\
&= \gamma_n \exp\left((\pi + \epsilon)\, \Gamma_n\right) \sum_{k \geq n} \gamma_k e^{-\epsilon \Gamma_k} \\
&\leq \gamma_n e^{(\epsilon - \pi)\Gamma_n} \int_{\Gamma_{n-1}}^{\infty} e^{-\epsilon x} dx \\
&\leq \gamma_n \frac{e^{\epsilon \gamma_n}}{\epsilon} e^{\pi \Gamma_n}.
\end{aligned}$$

Hence $\sum_k \gamma_k^2 e^{\pi \Gamma_k} < \infty$, and furthermore for some constant $C > 0$

$$\sum_{k \geq n} \gamma_k^2 e^{\pi \Gamma_k} \leq C \gamma_n e^{\pi \Gamma_n}.$$

So, since we have assumed that $\Phi_{\gamma,n}^2$ converges, by Proposition 2.12, on the set $\{Y_\infty = 0\}$,

$$(1 - X_n) \leq C \theta_n \gamma_n e^{\Gamma'_n}$$

So on $\{Y_\infty = 0\}\, (\subset \{X_\infty = 1\})$

$$\begin{aligned}
\frac{(1 - X_n)}{\gamma_n} &\leq C \theta_n e^{\Gamma'_n} \\
&= C e^{\Gamma'_n} \prod_{k=1}^{n} \left(1 - \gamma_k \psi_k X_{k-1}\right) \\
&= C e^{\Gamma'_n} \xi \prod_{k=1}^{n} \left(1 - \gamma_k \psi_k\right) \\
&\leq C
\end{aligned}$$

almost surely. Hence $\sup_n \frac{1-X_n}{\gamma_n} < \infty$ almost surely. Now, since (**S3'**) $\Longrightarrow$ (**S3**) the second part of the proposition follows from Lemma 2.16. ■

Finally we can bring all the results together in this section and prove Theorem 2.4:

Proof of Theorem 2.4. Assume (**S1**), (**E1**) and (**E3**) hold.

Part 1 follows immediately from Proposition 2.11 and the assumption (**E3**).

That $(1 - X_n) \sim e^{-p_2\Gamma_n}$ almost surely on $\{Y_\infty = 0\}$ under the additional assumption (**S3'**) follows from Proposition 2.15 part 2 together with the assumption (**E3**) and Corollary 2.14.

It is also shown in the proof of Proposition 2.15 that

$$(\mathbf{S3'}) \implies \sum_{n\geq 1} \gamma_n^2 e^{\Gamma_n} < \infty$$

and so the result that $\mathbb{P}(X_\infty = 1, Y_\infty > 0) > 0$ under (**S3'**) follows from Proposition 2.12. Furthermore, since

$$\sum_{n\geq 1} \gamma_n^2 e^{\Gamma_n} < \infty \implies \sum_{n\geq 1} \gamma_n^2 < \infty$$

it follows from Proposition 2.13 that

$$\mathbb{P}(Y_\infty = 0) > 0 \iff \sum_{k\geq 1} \prod_{k=1}^{n} (1 - p_1\gamma_k) < \infty.$$

Finally, that $(1 - X_n) \sim e^{p_1\Gamma_n}$ almost surely on $\{X_\infty = 1\}$ under the additional assumptions (**S4**) and (**S5**) follows from Proposition 2.15 part 1 together with the assumption (**E3**) and Corollary 2.14. ■

2.7 Discussion of the Assumptions

To assess the generality of the results obtained here it is worth considering assumptions (**E1**) and (**E3**) in more detail. We focus on describing the implications for learning sequences of the form:

$$\gamma_n = \left(\frac{D}{D+n}\right)^\alpha$$

where $\alpha \in (0, 1], D \in (0, \infty)$ as has been standard in recent literature.

2.7.1 Assumption (E1)

This assumption is not very complicated, and quite mild.

Consider the case when the payoffs are generated by i.i.d. Bernoulli random variables. Then we can take

$$\phi(n) = \left[\frac{n}{2}\log\log\left(\frac{n}{4}\right)\right]^{\frac{1}{2}+\epsilon}$$

for any value of $\epsilon > 0$. So in this case $\limsup_{n\to\infty} \gamma_n \phi(n) < \infty$ is true for all learning sequences

$$\gamma_n = \left(\frac{D}{D+n}\right)^{\alpha}$$

with $\alpha > \frac{1}{2}$.

In the worst case when we can only take $\phi(n) = n$ the assumption will still hold whenever $\alpha \geq 1$.

2.7.2 Assumption (E3)

In their paper Tarrès and Vandekerkhove [7] set out conditions under which $\Phi^a_{n,\xi}$ converges to a finite limit:

Lemma 2.17 *Given a positive real-valued non-increasing sequence $(\xi_n)_{n\geq 1}$ let, for all n, $\Xi_n := \sum_{k=1}^{n} \xi_k$. Suppose that ϕ is non-decreasing on $[k_0, \infty)$ for some $k_0 \in \mathbb{N}$, and if*

$$\sum_{k=1}^{\infty} \Xi_k \left|\phi''(k)\right| < \infty \text{ and } \limsup_{n\in\mathbb{N}} \Xi_n \left|\phi'(n)\right| = 0$$

where

$$\phi'(k) := \phi(k) - \phi(k-1) \text{ and } \phi''(k) := \phi'(k+1) - \phi'(k).$$

Then $\Phi^a_{n,\xi}$ converges to a finite real value as $n \to \infty$, for each $a = 1, 2$.

Suppose our sequence γ_n is given to us in a form $\gamma_n = g(n)$ for some decreasing function $g \in C([1,\infty))$. Then

$$\Gamma_n = \sum_{k=1}^{n} \gamma_k = O\left(\int_1^n g(x)\,dx\right).$$

From this we can estimate Γ_n for the two families of learning sequences we considered above:

$$\left(g(n) = \frac{D}{D+n},\ D > 0\right) \implies \Gamma_n = O(\log(n))$$

$$\left(g(n) = \left(\frac{D}{D+n}\right)^{\alpha},\ \alpha \in (0,1)\right) \implies \Gamma_n = O\left(n^{1-\alpha}\right)$$

Now if we take

$$\phi(n) = \left[\frac{n}{2}\log\log\left(\frac{n}{4}\right)\right]^{\frac{1}{2}+\epsilon}$$

then we have

$$\begin{aligned}\phi'(n) \sim &\left(\frac{1}{2}+\epsilon\right)\frac{\log\log\left(\frac{n}{4}\right)}{\left[n\log\log\left(\frac{n}{4}\right)\right]^{\frac{1}{2}-\epsilon}} \\ &+\left(\frac{1}{2}+\epsilon\right)\frac{1}{\log\left(\frac{n}{4}\right)\left[n\log\log\left(\frac{n}{4}\right)\right]^{\frac{1}{2}-\epsilon}}\end{aligned}$$

and

$$\begin{aligned}|\phi''(n)| \sim \Bigg|\frac{1}{n^{1-2\epsilon}}\Bigg[&\frac{\left(\frac{1}{2}+\epsilon\right)}{n^{\frac{1}{2}+\epsilon}\log\left(\frac{n}{4}\right)\left(\log\log\left(\frac{n}{4}\right)\right)^{\frac{1}{2}+\epsilon}} \\ &-\frac{\left(\frac{1}{2}-\epsilon\right)\left(\log\log\left(\frac{n}{4}\right)\right)^{\frac{1}{2}+\epsilon}}{n^{\frac{1}{2}+\epsilon}}\Bigg]\Bigg|.\end{aligned}$$

So in the case of i.i.d. Bernoulli random payoffs, the conditions of the Lemma hold for all $\alpha \in \left(\frac{1}{2}, 1\right]$.

Furthermore if we take $\phi(n) = n(\log n)^{-(1+\epsilon)}$ as in assumption (**E2**) then

$$\phi'(n) = (\log n)^{-(1+\epsilon)} - (1+\epsilon)(\log n)^{-(2+\epsilon)}$$

and

$$\phi''(n) = \frac{1+\epsilon}{n}\left[(\log n)^{-(1+\epsilon)} + (2+\epsilon)(\log n)^{-(3+\epsilon)}\right].$$

So in this case the conditions of the Lemma can only be satisfied when $\alpha = 1$.

However if we take can only take $\phi(n) = n$, then we have

$$\sum_{n=1}^{N} (p_a - \eta_{a,n}) \cdot \frac{1}{n} = \frac{\sum_{n=1}^{N} (p - \eta_{a,n})}{N} + \sum_{n=1}^{N} \left[\frac{1}{n-1} \frac{\sum_{k=1}^{n} (p - \eta_{a,k})}{n} \right]$$

and so it is uncertain whether or not $\Phi_{n,\xi}^{a}$ diverges when $\alpha = 1$. Of course $\Phi_{n,\xi}^{a}$ will converge when $\alpha > 1$, but then $\sum_{n \geq 0} \gamma_n < \infty$.

Remark 2.18 *Note that $\sum \gamma_k^2 < \infty$ does not imply that $\sum_{k=1}^{n} \gamma_k (\eta_{a,k} - p_a)$ converges when we can take $\phi(n) = \frac{n}{(\log(n+2))^{1+\epsilon}}$ ($g(n) = \frac{1}{n^{\alpha}}$ fails for all $\alpha \in \left(\frac{1}{2}, 1\right)$). Tarrès and Vandekerkhove do prove that the conditions are satisfied when*

$$\gamma_n = O\left(\Gamma_n e^{-p_2 \Gamma_n}\right) \text{ and we can take } \phi(n) = \frac{n}{(\log(n+2))^{1+\epsilon}}.$$

The key result in this proof is that $\gamma_n = O\left(\Gamma_n e^{-p_2 \Gamma_n}\right)$ implies $\Gamma_n = O(\ln(n))$. It is also true that although

$$\gamma_n = O\left(\Gamma_n e^{-p_2 \Gamma_n}\right) \implies \sum \gamma_k^2 < \infty$$

the converse does not hold.

2.7.3 The case $\phi \equiv Id$

Suppose that we cannot take a stronger ergodic condition than $\phi(n) = n$, and consider the step sequence

$$\left(\frac{D}{D+n} \right)_{n \geq 0}.$$

Now

$$\sum_{k=1}^{n} \eta_{2,k} \frac{D}{D+k} = \frac{D}{D+n} \sum_{k=1}^{n} \eta_{2,k} + \sum_{k=1}^{n-1} \frac{D}{(D+k)(D+k+1)} \sum_{i=0}^{k} \eta_{2,i}$$

and so

$$\frac{\sum_{k=1}^{n} \eta_{2,k} \frac{D}{D+k}}{\ln n^{Dp_2}} \to 1 \text{ as } n \to \infty.$$

Hence, since $\sum_{k\geq 1}\gamma_k^2 < \infty$,

$$\left(\sum_{n\geq 1}\prod_{k=1}^{n}(1-\eta_{2,k}\gamma_k)=\infty\right) \iff D \geq \frac{1}{p_2}$$

Now considering the proof of Lemma 2.8 we find that, when $\gamma_n = \frac{D}{D+n}$ we do not need the assumption (**E3**). Thus when $\alpha = 1$, the algorithm is consistent iff $D < \frac{1}{p_2}$. We will discuss the practical choice of D in the conclusion.

Furthermore it is possible by considering just Proposition 2.11 to show that the rate of convergence on $\{X_n \to 1\}$ at worst

$$n^{-D(p_1-p_2)}.$$

2.8 Conclusion

We have seen that many of the arguments in [2] can be well adapted from the annealed to the quenched context. We have achieved generalisations of Lamberton and Pagès' characterisation of consistency and rate of convergence results. From these results we can produce a general summary of the behaviour within the family of step sequences

$$\left\{\left(\left(\frac{D}{D+n}\right)^{\alpha}\right)_{n\geq 0}\right\}_{D,\alpha>0}$$

in analogy with that given in [2]:

Suppose that (**E1**) and (**E3**) hold. Then

- the strategy is inconsistent if

$$\left(\alpha \in \left(\frac{1}{2},1\right)\right) \text{ or if } \left(\alpha = 1 \text{ and } D > \frac{1}{p_2}\right);$$

- the strategy is consistent if

$$\alpha = 1 \text{ and } D \leq \frac{1}{p_2};$$

- if $\alpha = 1$ and $D \geq \frac{1}{(p_1 - p_2)}$, then the rate of convergence on $\{X_n \to 1\}$ is

$$n^{-Dp_1};$$

- if $\alpha = 1$ and $\frac{1}{p_1} < D < \frac{1}{(p_1 - p_2)}$, then *two* rates of convergence occur with positive probability on $\{X_n \to 1\}$,

$$n^{-Dp_1} \text{ and } n^{-D\pi};$$

- if $\alpha = 1$ and $D \leq \frac{1}{p_1}$ then the rate of convergence on $\{X_n \to 1\}$ is

$$n^{-D\pi}.$$

From the discussion of the assumptions (**E1**) and (**E3**) in the previous section we find that we can at best expect them to hold for $\alpha \in \left(\frac{1}{2}, 1\right]$, with this full range achieved in the case of random i.i.d. Bernoulli payoffs.

However in the case when we have no information about the rate of convergence of the Cesaro sums of the payoff sequences we cannot rely on (**E3**). In this case $\phi(n) = n$ so that (**E1**) will only be satisfied for $\alpha = 1$ and (**E3.1**) or (**E3.2**) would only be satisfied on $\alpha \in (1, \infty)$. Yet by directly checking the conditions of the results we find when $\alpha = 1$ the strategy is still consistent iff $D < \frac{1}{p_2}$. Furthermore it is possible, by considering just Proposition 2.11 to show that the rate of convergence is at worst $n^{-D\pi}$.

In conclusion, this shows that in a quenched context the only sensible choice of step sequence for application is once again the choice $\alpha = D = 1$, and this choice yields a rate of convergence of at worst $n^{-\pi}$ which becomes extremely slow when p_1 and p_2 are close. However this is exactly the same as in the annealed context, thus demonstrating a great deal of robustness of the algorithm.

In further work we hope to perform a similar analysis of the Penalized Bandit (see [3]), a more complex variation of the Narendra algorithm with much improved convergence behaviour, achieving a rate of convergence of $n^{-\frac{1}{2}}$ irrespective of the distance between p_1 and p_2.

Bibliography

[1] Evan Fisher. On the law of the iterated logarithm for martingales. *Annals of Probability*, 20:675 – 680, 1992.

[2] D. Lamberton and G. Pagès. How fast is the bandit. *Stochastic Analysis and Applications*, 26:603–623, 2008.

[3] D. Lamberton and G. Pagès. A penalized bandit algorithm. *Electronic Journal of Probability*, 13:341–373, 2008.

[4] D. Lamberton, G. Pagès, and P. Tarrès. When can the two-armed bandit algorithm be trusted. *Annals of Applied Probability*, 14:1424–1454, 2004.

[5] K. S. Narendra and I.J. Shapiro. Use of stochastic automata for parameter self-optimization with multimodal performance criteria. *IEEE Transactions on Systems Science and Cybernetics*, 5:352 – 360, 1969.

[6] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400 – 407, 1951.

[7] P. Tarrès and P. Vandekerkhove. On ergodic two-armed bandits. Accepted to The Annals of Applied Probability (to appear 2011).

[8] D. Williams. *Probability with Martingales*. CUP, 1991.

Chapter 3

Some Asymptotic Aspects of the Behaviour of the Gittins Index Strategy for Bandits with Bernoulli Payoffs

Joint work with Pierre Tarrès

Abstract

We study asymptotic properties of the Gittins Index Strategy for Bernoulli Bandits. First we give a simple proof that results by Kelly [5] imply that strategy is inconsistent. Secondly, through studying Whittle's interpretation of the Gittins Index in terms of value functions [6] we prove new, explicit asymptotic upper and lower bounds for the Gittins Indices. In particular these bounds come close to giving the second term in the Taylor expansion of the indices for large numbers of observations. Thirdly in the case of 2-Armed Bernoulli Bandits where both actions have success probability less than a half we obtain a non-zero lower bound on the probability that the strategy chooses to implement only a suboptimal action infinitely often, and thus deduce a lower bound on the rate at which this probability goes to zero as the discount factor tends to one.

3.1 Introduction

In this chapter we study the celebrated Gittins Index Strategy that solves the problem of geometrically discounted reward optimisation for the multi-Armed Bernoulli Bandit. We ask how the strategy behaves asymptotically as time goes to infinity, in particular giving a proof of inconsistency and providing a non-zero lower bound on the rate at which the probability that the strategy fails to learn decays as the discount factor tends to 1. In order to achieve these results we develop new, explicit asymptotic upper and lower bounds for the Gittins Indices. All our results are limited to Bernoulli Bandit Problems. Our motivation comes from attempting to understand how questions of learning in Bandit Problems are related to questions of reward maximisation.

Model We deal with several objects in a Bernoulli Bandit problem:

- a set of N actions $A = \{1, \ldots, N\}$;
- a set of initial observations $I_0 \in \left(\{0,1\}^2\right)^N$;

- a random sequence of actions $(A_n)_{n\geq 1} \in A^{\mathbb{N}_0}$ and, for each action a, a sequence of random variables $(O_{a,n})_{n\geq n}$ taking values in $\{0,1\}$, together with a filtration, $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$, where

$$\mathcal{F}_n = \sigma\left(I_0, (A_k)_{1\leq k\leq n}, \left\{(O_{a,k})_{1\leq k\leq n}\right\}_{a\in A}\right);$$

- a (possibly random) transition mechanism $P : \mathbb{N}_0 \times A \to \mathcal{P}(\{0,1\})$[1] such that $P(n,\cdot)$ is $\mathcal{F}_n$-measurable;

- a partial history space

$$\mathcal{H} := \bigcup_{n\in\mathbb{N}_0} \left(\mathbb{N}_0^2\right)^N \times (A\times\{0,1\})^n$$

  together with the set, Π, of maps $\pi : \mathcal{H} \to \mathcal{P}(A)$ which we call strategies;

- a reward map $r : \mathcal{H} \to \mathbb{R}$.

Let

$$\mathcal{X}_n := (I_0, A_1, I_1, \ldots A_n, I_n)$$

where $I_n := O_{A_n,n}$ for all $n \geq 1$. Then given the transition mechanism P and a strategy π we specify the transition probabilities of the random sequences $(A_n)_{n\geq 1}$ and $(O_{a,n})_{n\geq 1}$, $a \in A$, by, for all $n \in \mathbb{N}_0$ and $a \in A$,

$$\mathbb{P}(A_{n+1} = a \mid \mathcal{F}_n) := \pi(\mathcal{X}_n)(a)$$

and

$$\mathbb{P}(O_{a,n+1} = 1 \mid \mathcal{F}_n) := P(n,a)(1).$$

Thus $(\mathcal{X}_n)_{n\geq 0}$ is the random process of partial histories with $\mathcal{X}_0 = I_0$ and transition probabilities

$$\mathbb{P}(A_{n+1} = a,\ I_{n+1} = 1 \mid \mathcal{F}_n) = \pi(\mathcal{X}_n)(a)\, P(n,a)(1).$$

We call the triple (A, P, r) an N-Armed Bernoulli Bandit.

In a Bernoulli Bandit Problem an agent is presented with an N-Armed Bernoulli Bandit about which the set A, and the map r are made known to

[1] In this chapter for any set S, $\mathcal{P}(S)$ denotes the set of probability distributions over S

him, but he is given only imperfect information about P. The agent is then asked to choose a strategy $\pi \in \Pi$ such that some functional of the random reward sequence $\left(r\left(\mathcal{X}_n\right)\right)_{n\geq 1}$ has a certain desired property.

Write $I_0 = \left\{\left(s_{a,0}, f_{a,0}\right)\right\}_{a\in A}$ and let, for all $n \geq 1$,

$$s_{a,n} := s_{a,0} + \sum_{n\geq k\geq 1} \mathbf{1}_{\{A_k=a, I_k=1\}},$$
$$f_{a,n} := f_{a,0} + \sum_{n\geq k\geq 1} \mathbf{1}_{\{A_k=a, I_k=0\}}.$$

Then $s_{a,n}$, respectively $f_{a,n}$, represents the total number of success, respectively failures, observed by the agent from action a at time n.

We shall work with two separate assumptions on the transition mechanism:

A1 Define for each action a an additional sequence of random variables $\left(p_{a,n}\right)_{n\geq 0}$ such that $p_{a,n} \sim Beta\left(s_{a,n}+1, f_{a,n}+1\right)$ and extend the filtration $\mathcal{F}$ so that each $p_{a,n}$ is $\mathcal{F}_n$-measurable. Then we let

$$P\left(n,a\right)\left(1\right) := p_{a,n} \text{ for all } a \in A,\, n \in \mathbb{N}_0.$$

A2 For each $a \in A$, $P\left(n,a\right)\left(1\right) := p_a \in [0,1]$ for all $n \geq 0$ (so that $P\left(n,a\right)\left(1\right)$ is not random).

We can interpret A1, respectively A2, as a Bayesian, respectively frequentist, interpretation of the transition mechanism of an N-Armed Bandit. We shall use A1 only to define the Gittins Index Strategy, while we shall perform our analysis of the strategy under A2.

We normally assume that, for $n \geq 1$,

$$r\left(\mathcal{X}_n\right) := I_n (= O_{A_n,n}). \tag{3.1}$$

We use the general definition only when we come to define the Gittins Indices in Section 3.2.

Bayesian Reward Optimisation A1 corresponds to a decision by the agent to make the modelling assumption that each sequence $\left(O_{a,n}\right)_{n\geq 1}$ is an i.i.d. Bernoulli sequence with an unknown parameter; then $p_{a,0}$ can be interpreted as his choice of prior distribution for this parameter given the

initial information I_0, and $p_{a,n}$, $n \geq 1$, can be interpreted as the posterior distribution derived from the prior $p_{a,0}$ and the observations $\mathcal{X}_n$.

Let $\alpha \in (0,1)$. The *total expected α-discounted reward* under the strategy π from time n is defined to be

$$V_{\alpha,\pi}\left(\mathcal{X}_n\right) := \mathbb{E}_\pi\left(\sum_{k=1}^{\infty} \alpha^{k-1} r\left(\mathcal{X}_{n+k}\right) \mid \mathcal{X}_n\right),$$

and we let

$$V_\alpha\left(\mathcal{X}_n\right) := \sup_{\pi \in \Pi} V_{\alpha,\pi}\left(\mathcal{X}_n\right).$$

If a strategy π^* satisfies

$$V_{\alpha,\pi^*}\left(\mathcal{X}_0\right) = V_\alpha\left(\mathcal{X}_0\right)$$

for all initial observations $\mathcal{X}_0$ it is said to be *α-optimal*.

This leads to the following question of reward optimisation: Can we find a strategy that is α-optimal under assumption A1?

The Gittins Index Strategy In 1974 Gittins and Jones detailed a strategy, the Gittins Index Strategy, denoted here by π_G, which solves this problem of reward optimisation [1]. The strategy operates by first calculating, at each time n and for each action a, a number, or *index*, that depends only on the pair $(s_{a,n-1}, f_{a,n-1})$, so that the Index is a map

$$\nu_\alpha : \mathbb{N}_0^2 \to \mathbb{R}.$$

The strategy then chooses to implement an action with the highest index at that time. (See Section 3.2.1 for the definition of ν_α).

The Gittins Index Strategy has been found to be optimal in many more general settings, for example settings where the sequences $(O_{a,n})_{n\geq 1}$ are not Bernoulli but Normal [2], or settings in continuous time [4, 3]. However the computation of the indices is itself a difficult problem; normally they are approximated through some iterative procedure which becomes computationally intensive as the time until which one requires to play the strategy gets large (if one wishes to play for an arbitrary amount of time one may need to approximate the indices to an arbitrary degree of accuracy).

Bandit Problems and Learning We say that a strategy π is *consistent* if under the assumption A2

$$\mathbb{P}_\pi \left(\frac{1}{n} \sum_{i=0}^{n-1} r(\mathcal{X}_{i+1}) \to \max_{a \in A} \{p_a\} \text{ as } n \to \infty \right) = 1$$

and we say it is *inconsistent* otherwise.

Note that the definition of the Gittins Index Strategy π_G, and indeed of all possible strategies, is independent of whether assumption A1 or assumption A2 holds. Therefore we may certainly analyse the behaviour of the strategy π_G under the assumption A2 and ask whether it is consistent.

Asymptotic Behaviour of π_G In [5] Kelly provided an asymptotic analysis of the Gittins Index Strategy as $\alpha \to 1$. Along with his main results he showed that ν_α satisfies the following Lemma:

Lemma 3.1 *For all $\alpha \in (0,1)$ and $(s,f) \in \mathbb{N}_0^2$,*

$$\nu_\alpha (s,f) \geq \frac{s+1}{s+f+2} \tag{3.2}$$

Let S_n be a simple random walk on $\mathbb{Z}$, with $S_0 = 0$ and $\mathbb{P}(S_{n+1} - S_n = 1) = 1 - \mathbb{P}(S_{n+1} - S_n = -1)$ for $n \geq 1$. Then

$$\nu_\alpha (S_n, n - S_n) - \frac{S_n + 1}{n+2} \to 0 \text{ as } n \to \infty, \text{ almost surely.} \tag{3.3}$$

Using this and other results on implicit upper and lower bounds for ν_α Kelly characterised the behaviour of the Gittins Index strategy. He showed that, as $\alpha \to 1$, π_G tends to the strategy of Least Failures, π_{LF}, which is defined as follows:

$$\pi_{LF}(\mathcal{X}_n) \left(\arg\min_{a \in A} \{f_{a,n}\} \right) = 1.$$

Furthermore he showed that, for fixed α, the strategy only implements one action infinitely often, almost surely (this follows easily from Lemma 3.1).

Another well known result about the index (see for example [6]) states that, for any $(s, f) \in \mathbb{N}_0^2$,

$$\nu(s, f) \leq \nu(s+1, f) \text{ and } \nu(s, f) \geq \nu(s, f+1). \tag{3.4}$$

For the case when $N = 2$ we show in Section 3.2.2 an elementary proof of the folk result that π_G is also inconsistent using (3.4) and Lemma 3.1, i.e. that the action the strategy eventually implements for ever is suboptimal with non-zero probability. However in order to put a positive lower bound on the probability that π_G implements the suboptimal action infinitely often it is desirable to have not just an explicit lower bound, but also an explicit upper bound for ν_α.

Bounds for ν_α and the Inconsistency of π_G Recall that ν_α is defined on the set $\mathbb{N}_0^2$. Consider the following change of coordinates

$$(s, f) \mapsto \left(x, \frac{1}{n}\right) := \left(\frac{s+1}{s+f+2}, \frac{1}{s+f+2}\right)$$

and the corresponding change of domain

$$\mathbb{N}_0^2 \mapsto \mathbb{G} := \left\{\left(\frac{s+1}{s+f+2}, \frac{1}{s+f+2}\right) : (s, f) \in \mathbb{N}_0^2\right\}$$

Then we can consider the Gittins Index as a function $\nu_\alpha : \mathbb{G} \to \mathbb{R}$.

In Section 3.3 we derive the following new explicit bounds for the Gittins Index in these new coordinates:

Theorem 3.2 (Explicit Asymptotic Bounds for the Gittins Index) *Fix $\alpha \in (0, 1)$.*

1. *There is an asymptotic upper bound $\nu_U : \mathbb{G} \to \mathbb{R}$ for the Gittins Index given by*

$$\nu_U\left(x, \frac{1}{n}\right) = x + \frac{\alpha}{1-\alpha} x(1-x)\frac{1}{n}$$

and it is valid when $\frac{1}{n} \leq \frac{1-\alpha}{\alpha}$.

2. *There is an asymptotic lower bound* $\nu_L : \mathbb{G} \to \mathbb{R}$ *for the Gittins Index given by*

$$\nu_L\left(x, \frac{1}{n}\right) = x + \frac{\alpha}{1-\alpha x} x\left(1-x\right)\frac{1}{n} + O\left(\frac{1}{n^2}\right).$$

We note that this Theorem comes close to giving the second term in the Taylor expansion of the index ν_α around the time $n = \infty$, particularly for large values of x. If the bounds could be refined to give this second term it could be used to calculate the Gittins Indices more efficiently.

Under the additional assumptions that $|A| = 2$ and that $p_2 \leq \frac{1}{2}$ and that $p_1 > p_2$, using the upper bound in Theorem 3.2 we are able to prove a preliminary lower bound for the rate of decay of the probability that the strategy fails to learn as $\alpha \to 1$:

Proposition 3.3 *Assume that A2 holds, that* $|A| = 2$*, that* $p_2 \leq \frac{1}{2}$ *and that* $p_1 > p_2$*. If*

$$E := \{A_n = 1 \text{ finitely often}\}$$

then for any fixed $\epsilon > 0$

$$\lim_{\alpha \to 1} \mathbb{P}\left(E\right).\left(1-p_1\right)^{\frac{\alpha}{1-\alpha}} \geq C\left(p_2, \epsilon\right) > 0$$

where $C\left(p_2, \epsilon\right)$ *is a constant for fixed* ϵ*.*

The rest of the chapter is structured as follows. In Section 3.2 we give a definition of the index ν_α and an elementary proof of the inconsistency of the strategy π_G. In Section 3.3 we derive the new, explicit, asymptotic bounds for ν_α given in Theorem 3.2. Finally in Section 3.4 we prove a non-zero lower bound for the probability that the π_G fails to learn from which we deduce Proposition 3.3.

3.2 The Gittins Index and Inconsistency of π_G

3.2.1 The Gittins Indices

There are various descriptions of the Gittins Indices, but we choose to reproduce Whittle's interpretation here [6], as it is the key to our explicit bounds on the Index which we obtain in Section 3.3.

Suppose that A1 and (3.1) hold, and recall the function

$$V_\alpha(\mathcal{X}_n) := \sup_{\pi \in \Pi} V_{\alpha,\pi}(\mathcal{X}_n), \ \forall \mathcal{X}_n \in \mathcal{H}.$$

We call this function the *value function* and note that it is the unique, bounded solution of the dynamic optimality equation [6]:

$$V_\alpha(\mathcal{X}_n) = \max_{a \in A} L_a V_\alpha(\mathcal{X}_n), \ \forall \mathcal{X}_n \in \mathcal{H},$$

where L_a is the one step function:

$$\begin{aligned} L_a V_\alpha(\mathcal{X}_n) := \mathbb{E}(r(\mathcal{X}_{n+1}) \mid \mathcal{X}_n, A_{n+1} = a) \\ + \alpha \mathbb{E}(V_\alpha(\mathcal{X}_{n+1}) \mid \mathcal{X}_n, A_{n+1} = a). \end{aligned}$$

Now suppose that $N = 2$ and consider a modification of the reward function so that

$$r(\mathcal{X}_n) := \begin{cases} I_n & A_k = 1 \text{ for all } k \in \{1, \ldots, n\}, \\ M & A_k = 1 \text{ for all } k \in \{1, \ldots, n-1\} \text{ and } A_n = 2, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where M is some constant in the interval $\left[0, \frac{1}{1-\alpha}\right]$. We call a 2-Armed Bernoulli Bandit with reward function satisfying (3.5) a 1-*Armed Bernoulli Bandit with retirement value M*.

The value function of this bandit, denoted by V_α^M, will also be the unique bounded solution of the dynamic optimality equation. Moreover by the construction of the reward function, r, the values of $s_{2,0}$ and $f_{2,0}$ do not influence the value of $V_\alpha^M(\mathcal{X}_0)$. We therefore, in an abuse of notation, consider V_α^M as a function supported on $\mathbb{N}_0^2$ which satisfies:

$$V_\alpha^M(s,f) = \max\left\{M, L_1 V_\alpha^M(s,f)\right\}, \ \forall (s,f) \in \mathbb{N}_0^2.$$

Clearly V_α^M is non-decreasing and continuous in M.

Using V_α^M we can define the Gittins Index:

Definition 3.4 (The Gittins Index) *Let V_α^M be the value function for a 1-Armed Bernoulli Bandit with retirement value M under the assumption A1. Then we define the* Gittins Index *to be the map $\nu_\alpha : \mathbb{N}_0^2 \to \mathbb{R}$ given by*

$$\nu_\alpha(s,f) := (1-\alpha) \inf\left\{M \in \mathbb{R} : V_\alpha^M(s,f) = M\right\}$$

for all $(s,f) \in \mathbb{N}_0^2$.

Notation 3.5 *Although V_α, V_α^M, and ν_α depend on the parameter of the discount sequence, α, we often suppress this in the notation when it is not important, instead writing just V, V^M, or ν respectively.*

This definition from Whittle affords us the intuitive interpretation of the Gittins Index of an action, given the initial information and history of observations made from that action, as the "normalised retirement value" of that action.

As a further illustration of this definition we give a proof of the lower bound for the index given in Lemma 3.1:

Proof of (3.2) in Lemma 3.1. Consider V^M, the value function of a 1-Armed Bernoulli Bandit with retirement value M under assumption A1.

Write $\mathcal{X}_0 = (s, f) \in \mathbb{N}_0^2$. Now since V^M is continuous in M we can choose $M_{s,f}$ so that

$$(1-\alpha) M_{s,f} = \nu_\alpha (s, f).$$

Then

$$\begin{aligned} M_{s,f} &= V^{M_{s,f}} (\mathcal{X}_0) \\ &= \mathbb{E} (O_{1,1} \mid \mathcal{X}_0) + \alpha \mathbb{E} \left(V^{M_{s,f}} (\mathcal{X}_1) \mid \mathcal{X}_0, A_1 = 1 \right) \end{aligned} \tag{3.6}$$

But it follows from the definition of r that when $A_1 = 1$, $V^{M_{s,f}} (\mathcal{X}_1) \geq M_{s,f}$. So

$$\mathbb{E} \left(V^{M_{s,f}} (\mathcal{X}_1) \mid \mathcal{X}_0, A_1 = 1 \right) \geq M_{s,f}$$

and plugging this into (3.6) we obtain

$$M_{s,f} \geq \mathbb{E} (O_{1,1} \mid \mathcal{X}_0) + \alpha M_{s,f}.$$

Since under A1

$$\mathbb{E} (O_{1,1} \mid \mathcal{X}_0) = \frac{s+1}{s+f+2}$$

(3.2) follows immediately. ■

Remark 3.6 *The second statement in Lemma 3.1, that (3.3) holds under the assumption A2, can also be proved in this language, for example as a corollary of Lemma 3.10 below.*

3.2.2 Inconsistency of the Gittins Index Strategy

Using Lemma 3.1 we can prove the folk result that the Gittins Index Strategy for 2-Armed Bernoulli Bandits when the sequences $(O_{a,n})_{n\geq 1}$, $a \in A$, satisfy A2 is inconsistent:

Theorem 3.7 *Consider a 2-Armed Bernoulli Bandit under assumption A2, and suppose that*

$$1 > p_1 \neq p_2 > 0.$$

Then the Gittins Index Strategy is inconsistent.

Remark 3.8 *From Lemma 3.1 and the Strong Law of Large Numbers we know that under assumption A2 if the agent implements action $a \in A$ infinitely often, then $\nu(s_{a,n}, f_{a,n}) \to p_a$ as $n \to \infty$ almost surely. So as Kelly noted in [5] we therefore easily deduce by a contradiction that under the assumptions of Theorem 3.7 the strategy π_G implements only one action infinitely often almost surely. It remains only to show that the strategy chooses to implement the suboptimal action infinitely often with positive probability, and this is what we show in this proof.*

Proof. Suppose wlog that $p_1 > p_2$.

Let

$$x_{2,n} := \frac{s_{2,n} + 1}{s_{2,n} + f_{2,n} + 2}$$

and set $0 < \epsilon < p_2$. First we show that from Lemma 3.1 it follows that the event

$$E_{\epsilon,n_0} := \{\nu(s_{1,n_0}, f_{1,n_0}) < p_2 - \epsilon\} \cap \{x_{2,n_0} > p_2 - \epsilon\}$$

is not null for some n_0.

Indeed suppose that for all $0 \leq n \leq m_0$, $I_{n+1} = 0$. Then it follows from the second part of Lemma 3.1 together with (3.1) that for sufficiently large m_0

$$\begin{gathered}\nu(s_{1,m_0}, f_{1,m_0}) < \min\{p_2 - \epsilon, \nu(s_{2,0}, f_{2,0})\} \\ \text{and } \nu(s_{1,m_0}, f_{1,m_0}) \leq \nu(s_{2,m_0}, f_{2,m_0}).\end{gathered}$$

Supposing that $\nu(s_{1,m_0}, f_{1,m_0}) \leq \nu(s_{2,m_0}, f_{2,m_0})$, and that $A_{m_0+1} = 2$, then it follows from the first part of Lemma 3.1 and (3.1) that if $I_{n+1} = 1$ for all $m_0 < n \leq n_0$, then for sufficiently large n_0 we shall have

$$\nu(s_{1,n_0}, f_{1,n_0}) \leq \nu(s_{2,n_0}, f_{2,n_0}) \text{ and } x_{2,n_0} > p_2 - \epsilon.$$

Hence we find, by choosing m_0 and n_0 enough, that $\mathbb{P}(E_{\epsilon,n_0}) \geq (1-p_1)^{m_0} p_2^{n_0} > 0$, i.e. that E_{ϵ,n_0} is not null for some n_0.

For $n \geq n_0$, let

$$\bar{x}_{2,n}(n_0) := \frac{s_{2,n_0} + 1 + \sum_{k=n_0+1}^{n} O_{2,k}}{s_{2,n_0} + f_{2,n_0} + 2 + (n - n_0)}$$

and

$$T_{2,p_2-\epsilon}(n_0) := \max\{n \geq n_0 : \bar{x}_{2,n} \leq p_2 - \epsilon\}.$$

Let $\bar{F}_{\epsilon,t}(n_0) := \{T_{2,p_2-\epsilon}(n_0) = n_0 + t\}$. Now E_{ϵ,n_0} is $\mathcal{F}_{n_0-1}$-measurable, and the sequence $(O_{2,n})_{n \geq n_0}$ is independent of $\mathcal{F}_{n_0}$. So it follows from the Strong Law of Large Numbers, that for some t,

$$\mathbb{P}\left(\bar{F}_{\epsilon,t}(n_0) \mid E_{\epsilon,n_0}\right) > 0$$

and hence that $\bar{F}_{\epsilon,t}(n_0) \cap E_{\epsilon,n_0}$ is not null for some t.

Now for each $\omega \in \bar{F}_{\epsilon,t}(n_0) \cap E_{\epsilon,n_0}$ there exists a permutation $\sigma_\omega \in S_{t+1}$ such that for all $n \in \{1, \ldots, t\}$

$$\frac{s_{2,n_0} + 1 + \sum_{k=1}^{n} O_{2,n_0+\sigma(k)}}{s_{2,n_0} + f_{2,n_0} + 2 + n} \geq p_2 - \epsilon.$$

So define the event

$$\begin{aligned} F_{\epsilon,t}(n_0) := \big\{ & \omega \in \bar{F}_{\epsilon,t}(n_0) \cap E_{\epsilon,n_0} : \exists \omega_0 \in \bar{F}_{\epsilon,t}(n_0) \cap E_{\epsilon,n_0} \\ & \text{s.t. } O_{2,n_0+k}(\omega) = O_{2,n_0+\sigma_{\omega_0}(k)}(\omega_0),\ k = 1, \ldots, t \big\}. \end{aligned}$$

Clearly $\mathbb{P}(F_{\epsilon,t}(n_0)) \geq \frac{1}{(t+1)!}\mathbb{P}\left(\bar{F}_{\epsilon,t}(n_0) \cap E_{\epsilon,n_0}\right)$, and so $F_{\epsilon,t}(n_0)$ is not null for some t.

Now it follows from (3.2) that on the event $F_{\epsilon,t}(n_0)$ the Gittins Index Strategy will choose to implement action 2 for all $n \geq n_0$. Since

$$\mathbb{P}(F_{\epsilon,t}(n_0)) > 0$$

for some ϵ, n_0 and t this proves that the strategy eventually implements only the suboptimal action for ever with positive probability. Hence π_G is inconsistent. ■

The proof easily generalises to N-Armed Bernoulli Bandits.

3.3 Explicit Bounds for the Gittins Index

3.3.1 Scheme for Deriving the Bounds

Recall once again that the Gittins Index is defined in terms of the value function V^M of a 1-Armed Bernoulli Bandit with retirement value M under the assumption A1:

$$\nu_\alpha(s,f) := (1-\alpha)\inf\left\{M \in \mathbb{R} : V^M(s,f) = M\right\}.$$

We therefore make a study of this value function.

From the dynamic optimality equation we know that V^M satisfies, for all $(s,f) \in \mathbb{N}_0^2$,

$$V^M(s,f) = \max\left\{M, L_1V^M(s,f)\right\} \tag{3.7}$$

where

$$L_1V^M(s,f) = \frac{s+1}{s+f+2} + \alpha\left[\frac{s+1}{s+f+2}V^M(s+1,f) + \frac{f+1}{s+f+2}V^M(s,f+1)\right]$$

Now since we are interested in obtaining explicit bounds for the indices as $s+f \to \infty$ we make the change of coordinates introduced in Section 3.1

$$(s,f) \mapsto (x,t) = \left(\frac{s+1}{s+f+2}, \frac{1}{s+f+2}\right)$$

and the corresponding change of domain

$$\mathbb{N}_0^2 \mapsto \mathbb{G} := \bigcup_{n\geq 2}\mathbb{G}_n,$$

where

$$\mathbb{G}_n := \left\{\left(\frac{k}{n}, \frac{1}{n}\right) : 2 \leq n \in \mathbb{N},\ k \in \{1,\ldots,n-1\}\right\}.$$

In these new coordinates $V^M : \mathbb{G} \to \mathbb{R}$ and we have

$$L_1V^M\left(x, \frac{1}{n}\right) = x + \alpha\left[xV^M\left(r_{n+1}(x), \frac{1}{n+1}\right) + (1-x)V^M\left(l_{n+1}(x), \frac{1}{n+1}\right)\right]$$

where

$$l_{n+1}(x) := x - \frac{x}{n+1},\ r_{n+1}(x) := x + \frac{(1-x)}{n+1}.$$

We now construct some approximations for V^M as follows. Intuitively we expect that

$$\left\| V^M\left(x, \frac{1}{n}\right) - \max\left\{M, \frac{x}{1-\alpha}\right\}\right\|_\infty \to 0 \text{ as } n \to \infty.$$

So let V^{M,n_0} be defined by, for all $\left(x, \frac{1}{n}\right) \in \mathbb{G}$,

$$V^{M,n_0}\left(x, \frac{1}{n}\right) := \max\left\{M, \frac{x}{1-\alpha}\right\} \text{ when } n \geq n_0,$$
$$\text{and } V^{M,n_0}\left(x, \frac{1}{n}\right) = \max\left\{M, L_1 V^{M,n_0}\left(x, \frac{1}{n+1}\right)\right\} \text{ when } n < n_0. \quad (3.8)$$

We prove in Subsection 3.3.2 that $V^{M,n_0}\left(x, \frac{1}{n}\right) \to V^M\left(x, \frac{1}{n}\right)$ as $n_0 \to \infty$.

Now $V^M\left(x, \frac{1}{n}\right)$ is increasing in x. So define

$$x_{\max}(n, M, \alpha) := \max_{k \in \{1,\ldots n-1\}} \left\{\frac{k}{n} : L_1 V^M\left(\frac{k}{n}, \frac{1}{n}\right) \leq M\right\}.$$

Then for all $\left(x, \frac{1}{n}\right) \in \mathbb{G}$

$$x \leq x_{\max}(n, M, \alpha) \implies \nu_\alpha\left(x, \frac{1}{n}\right) \leq M(1-\alpha) \quad (3.9)$$

and

$$x > x_{\max}(n, M, \alpha) \implies \nu_\alpha\left(x, \frac{1}{n}\right) > M(1-\alpha), \quad (3.10)$$

where we use continuity of $V^M(x, 1/n)$ in M in the second implication.

Our strategy is therefore to upper and lower bound $x_{\max}(n, M, \alpha)$ through the approximations $\left(V^{M,n_0}\right)_{n_0 \geq 1}$ whose behaviour we can calculate explicitly. Through taking the limit as $n_0 \to \infty$ we will then obtain both upper and lower bounds for the Gittins Index.

3.3.2 Tools for Deriving the Bounds

Some Useful Notation and Extending to a Continuous Domain

First we note that we may extend the domain of the approximations $\left(V^{M,n_0}\right)_{n_0\geq 2}$. Define the set

$$\mathbb{F} := \bigcup_{n\in\mathbb{N}} \mathbb{F}_n.$$

where,

$$\mathbb{F}_n := \left[\frac{1}{n}, \frac{n-1}{n}\right] \times \left\{\frac{1}{n}\right\}.$$

Then we can define V^{M,n_0} exactly as in (3.8) except taking $\left(x, \frac{1}{n}\right) \in \mathbb{F}$. We show later in Lemma 3.10 that V^M may also be extended to this domain.

Second we note that we can decompose the iteration (3.7) into three operators:

$$V^M\left(., \frac{1}{n}\right) = (W \circ L \circ T_{n+1})\left(V^M\left(., \frac{1}{n+1}\right)\right)$$

where

$$T_{n+1} : C_0\left(\left[\frac{1}{n+1}, \frac{n}{n+1}\right]\right) \to C_0\left(\left[\frac{1}{n}, \frac{n-1}{n}\right]\right)$$
$$\& \; L, W : C_0(0,1) \to C_0(0,1)$$

are defined, for any $U \in C_0\left(\left[\frac{1}{n+1}, \frac{n}{n+1}\right]\right)$ and $V \in C_0(0,1)$, by

$$\begin{aligned} T_{n+1}(U(x)) &:= xU(r_{n+1}(x)) + (1-x)U(l_{n+1}(x)), \\ L(V(x)) &:= x + \alpha V(x), \\ W(V(x)) &:= \max\{M, V(x)\}. \end{aligned} \tag{3.11}$$

Notice that T_{n+1} and L are both linear and non-negative (i.e. $U, V \geq 0$ then $T_{n+1}(U), L(V) \geq 0$).

Finally for any function $V : \mathbb{R} \to \mathbb{R}$ let, for any $x, y \in \mathbb{R}$,

$$\Delta_{x,y}(V) := \frac{V(y) - V(x)}{y - x}.$$

Armed with this notation we make our analysis of V^M.

Properties of the Value Function and the Approximations

We begin with a Lemma about the operators T_n and L.

Lemma 3.9 *Let $n \geq 2$, and let $V : \left[\frac{1}{n}, \frac{n-1}{n}\right] \to \mathbb{R}$.*

If $V(\cdot)$ is continuous, convex and non-decreasing, then so are $T_n(V(\cdot))$ and $L(V(\cdot))$.

Furthermore, if $V(\cdot)$ is $\frac{1}{1-\alpha}$-Lipschitz, then so are $T_n(V(\cdot))$ and $L(V(\cdot))$.

Proof. This can be checked from the definitions of T_n and L by elementary calculations ■

Using this Lemma we can prove the first important result about the approximations $\left(V^{M,n_0}\right)_{n_0 \geq 2}$ and the function V^M.

Lemma 3.10 *We note the following properties of the approximations $\left(V^{M,n_0}\right)_{n_0 \geq 2}$ and the function V^M:*

1. *For all $n_0, n \geq 2$, $V^{M,n_0}\left(\cdot, \frac{1}{n}\right)$ is continuous, convex, non-decreasing, $\frac{1}{1-\alpha}$-Lipschitz and bounded above by $\frac{1}{1-\alpha}$.*
2. *For each $n \geq 2$, V^{M,n_0} converges uniformly to a limit $\bar{V}^M$ as $n_0 \to \infty$ on the set $\bar{\mathbb{F}}_n := \bigcup_{k \leq n} \mathbb{F}_k$. Moreover $\bar{V}^M|_{\mathbb{G}} \equiv V^M$.*
3. *For all $n \geq 2$, $\bar{V}^M\left(\cdot, \frac{1}{n}\right)$ is continuous, convex, non-decreasing, and $\frac{1}{1-\alpha}$-Lipschitz.*

Proof. That $V^{M,n_0}\left(\cdot, \frac{1}{n}\right)$ is continuous, and convex and non-decreasing follows immediately from (3.8) and Lemma 3.9.

Note that for all $n \geq n_0$ and $\frac{1}{n_0} \leq x < y \leq \frac{n_0-1}{n_0}$,

$$0 \leq \Delta_{y,x}\left(V^{M,n_0}\right)\left(\cdot, \frac{1}{n}\right) \leq \frac{1}{1-\alpha}.$$

So it follows from Lemma 3.9 $V^{M,n_0}\left(\cdot, \frac{1}{n}\right)$ is $\frac{1}{1-\alpha}$-Lipschitz for all $n \geq 2$.

Note that $V^{M,n_0}\left(\cdot, \frac{1}{n_0}\right)$ is bounded above by $\frac{1}{1-\alpha}$. Now if $V^{M,n_0}\left(\cdot, \frac{1}{n}\right)$ is bounded above by $\frac{1}{1-\alpha}$, then

$$\begin{aligned}
&(L \circ T_n)\left(V^{M,n_0}\left(x, \frac{1}{n-1}\right)\right) \\
&\quad = x + \alpha\left[xV\left(r_n(x), \frac{1}{n}\right) + (1-x)V\left(l_n(x), \frac{1}{n}\right)\right] \\
&\quad \leq x + \frac{\alpha}{1-\alpha}\left[x + 1 - x\right] \leq \frac{1}{1-\alpha},
\end{aligned}$$

and so, by an induction we conclude that $V^{M,n_0} \leq \frac{1}{1-\alpha}$. This completes the proof of part 1.

Let us now prove part 2: we first prove by descending induction that, if $U, V : \mathbb{F} \to \mathbb{R}$ both satisfy (3.7) on $\bar{\mathbb{F}}_{n_0} := \bigcup_{n \leq n_0} \mathbb{F}_n$, then for all $2 \leq k \leq n_0$,

$$\|U(1/k, .) - V(1/k, .)\|_\infty < \alpha^{n_0 - k}\|U(1/n_0, .) - V(1/n_0, .)\|_\infty.$$

Indeed, suppose that for $k \leq n_0$ and some $\delta > 0$

$$\|U - V\|_\infty < \delta\alpha^{n_0-k} \text{ on } \mathbb{F}_k.$$

Then for all $\left(x, \frac{1}{k-1}\right) \in \mathbb{F}_{k-1}$

$$\begin{aligned}
&\left|U\left(x, \frac{1}{k-1}\right) - V\left(x, \frac{1}{k-1}\right)\right| \\
&\quad \leq \alpha\left[x\left|U\left(r_n(x), \frac{1}{k}\right) - V\left(r_n(x), \frac{1}{k}\right)\right|\right. \\
&\qquad\qquad \left. + (1-x)\left|U\left(l_n(x), \frac{1}{k}\right) - V\left(l_n(x), \frac{1}{k}\right)\right|\right] \\
&\quad \leq \alpha\delta\alpha^{n_0-k} = \delta\alpha^{n_0-(k-1)}
\end{aligned}$$

But, from part 1 we know that

$$\frac{1}{1-\alpha} \geq V^{M,n_0} \geq M$$

and taking this as our base case completes the induction. It now follows that V^{M,n_0} converges uniformly on each set $\bar{\mathbb{F}}_n$ to a function $\bar{V}^M : \mathbb{F} \to \mathbb{R}$.

Moreover, $\bar{V}^M$ satisfies (3.7) on $\mathbb{G}$ and so, since V^M is the unique bounded function satisfying (3.7) on $\mathbb{G}$ [6] this completes the proof of part 2.

Finally part 3 follows easily from parts 1 and 2. ■

Remark 3.11 *From this Lemma we see that we may extend V^M to the domain $\mathbb{F}$ by the function $\bar{V}^M$ and we henceforth use V^M to mean both the function restricted to $\mathbb{G}$ and the extension to $\mathbb{F}$.*

Notation 3.12 (Critical Points) *In Subsection 3.3.1 we have already defined*

$$x_{\max}(n, M, \alpha) := \max_{k=1,\ldots,n-1} \left\{ \frac{k}{n} : V^M\left(\frac{k}{n}, \frac{1}{n}\right) = M \right\}.$$

We define further

$$x_{\sup}(n, M, \alpha) := \sup \left\{ x \in \left[\frac{1}{n}, \frac{n-1}{n}\right] : V^M\left(x, \frac{1}{n}\right) = M \right\},$$

$$x_{\sup}^{n_0}(n, M, \alpha) := \sup \left\{ x \in \left[\frac{1}{n}, \frac{n-1}{n}\right] : V^{M,n_0}\left(x, \frac{1}{n}\right) = M \right\},$$

and

$$x_{\max}^{n_0}(n, M, \alpha) := \max_{k=1,\ldots,n-1} \left\{ \frac{k}{n} : V^{M,n_0}\left(\frac{k}{n}, \frac{1}{n}\right) = M \right\}.$$

The lemma below is not needed in the proofs of the upper and lower bounds, however it allows us to draw accurate pictures of the value function V^M on a computer using the approximations V^{M,n_0}.

Lemma 3.13 *$x_{\sup}^{n_0}(n, M, \alpha) \to x_{\sup}(n, M, \alpha)$ as $n_0 \to \infty$.*

Hence for sufficiently large n_0

$$x_{\max}(n, M, \alpha) = x_{\max}^{n_0}(n, M, \alpha). \tag{3.12}$$

Proof. Note that, for any n, M, andα, when $x < y < x_{\sup}(n, M, \alpha)$

$$(L \circ T_{n+1})\left(V^M\left(x, \frac{1}{n+1}\right)\right) < (L \circ T_{n+1})\left(V^M\left(y, \frac{1}{n+1}\right)\right)$$

From this we deduce that whenever $x \in (0, x_{\max}(n, M, \alpha))$

$$(L \circ T_{n+1})\left(V^M\left(x, \frac{1}{n+1}\right)\right) < M.$$

Recall that, for each $n \geq 2$, by Lemma 3.10 the approximations $V^{M,n_0}\left(\cdot, \frac{1}{n}\right)$ are converging uniformly to $V^M\left(\cdot, \frac{1}{n}\right)$ on $\left[\frac{1}{n}, \frac{n-1}{n}\right]$ as $n_0 \to \infty$. Hence we deduce that $x_{\sup}^{n_0}(n, M, \alpha) \to x_{\sup}(n, M, \alpha)$ as $n_0 \to \infty$.

(3.12) now follows easily. ■

Using the approximation $V^{M,100}$ we plot the graph

$$\left\{\left(\frac{k}{n}, V^M\left(\frac{k}{n}, \frac{1}{n}\right)\right) : 1 \leq k \leq n-1\right\}$$

for $n = 40$, $M = 50$, $\alpha = 0.99$ (see figure 3.1). Using the approximation $V^{M,301}$ we plot the graph

$$\left\{\left(x^{301}_{\max}(n, M, \alpha), \frac{1}{n}\right) : 3 \leq n \leq 301\right\}$$

when $\alpha = 0.95$ and $M = \frac{15}{(1-\alpha).10\pi}$ (see figure 3.2).

Bounding the Approximations

To recap, our strategy is to analyse the value function V^M through the approximations V^{M,n_0}. Later in our analysis we will describe simple, explicit bounds for the $\frac{1}{n}$-slices of the approximations V^{M,n_0}, and use a limiting argument to show that these bounds also bound the $\frac{1}{n}$-slices of V^M.

To find these bounds we look for some stability properties of the operators $(W \circ L \circ T_n)$, $n > 2$, by considering how these operators act on the family of piecewise linear functions:

$$V^M_{m,\epsilon} : x \to M + \frac{m}{1-\alpha}(x - (b-\epsilon))_+,$$

where $x_+ := \max(x, 0)$ denotes the positive part of x, $m \in [0, 1]$, $\epsilon \in [0, b]$, and

$$b := M.(1-\alpha).$$

From this family of functions we shall be able to bound above and below the approximations $\left(V^{M,n_0}\right)_{n_0 \geq 0}$ and hence also V^M.

In this subsection we therefore study how $(W \circ L \circ T_n)$, and, in particular, how T_n and L individually act on the simple functions $V^M_{m,\epsilon} \in C_0(0, 1)$.

First consider T_n in isolation. It easy to see by inspection that

$$T_n\left(V^M_{m,\epsilon}(x)\right) \in C_1\left(I_{b,\epsilon,n}\right),$$

where $I_{b,\epsilon,n}$ is the interval

$$I_{b,\epsilon,n} := \left(\frac{n}{n-1}(b-\epsilon) - \frac{1}{n-1}, \frac{n}{n-1}(b-\epsilon)\right),$$

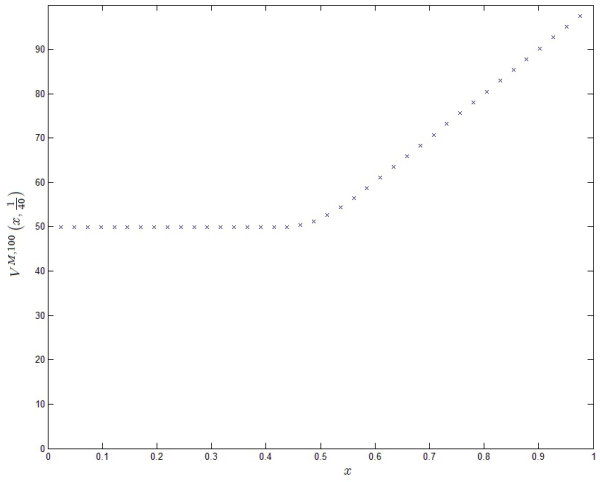


Figure 3.1: The values $V^{M,100}\left(x,\frac{1}{40}\right)$ plotted against x for all $\left(x,\frac{1}{40}\right) \in \mathbb{G}_{40}$ when $M = 50$ and $\alpha = 0.99$.

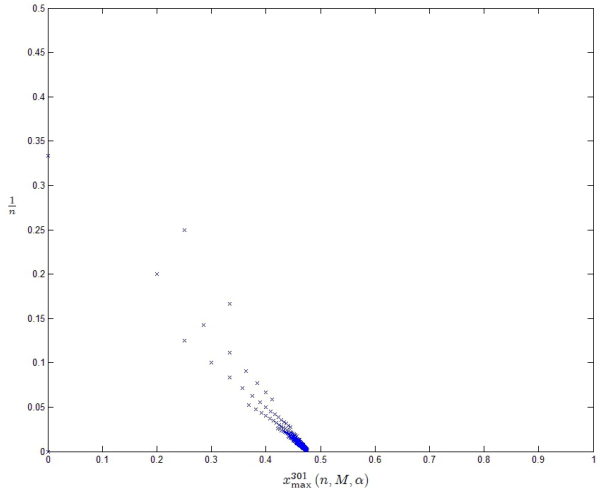


Figure 3.2: The graph $\left\{\left(x^{301}_{\max}\left(n,M,\alpha\right),\frac{1}{n}\right)\right\}_{301\geq n\geq 3}$ when $\alpha = 0.95$ and $M = \frac{15}{(1-\alpha).10\pi}$.

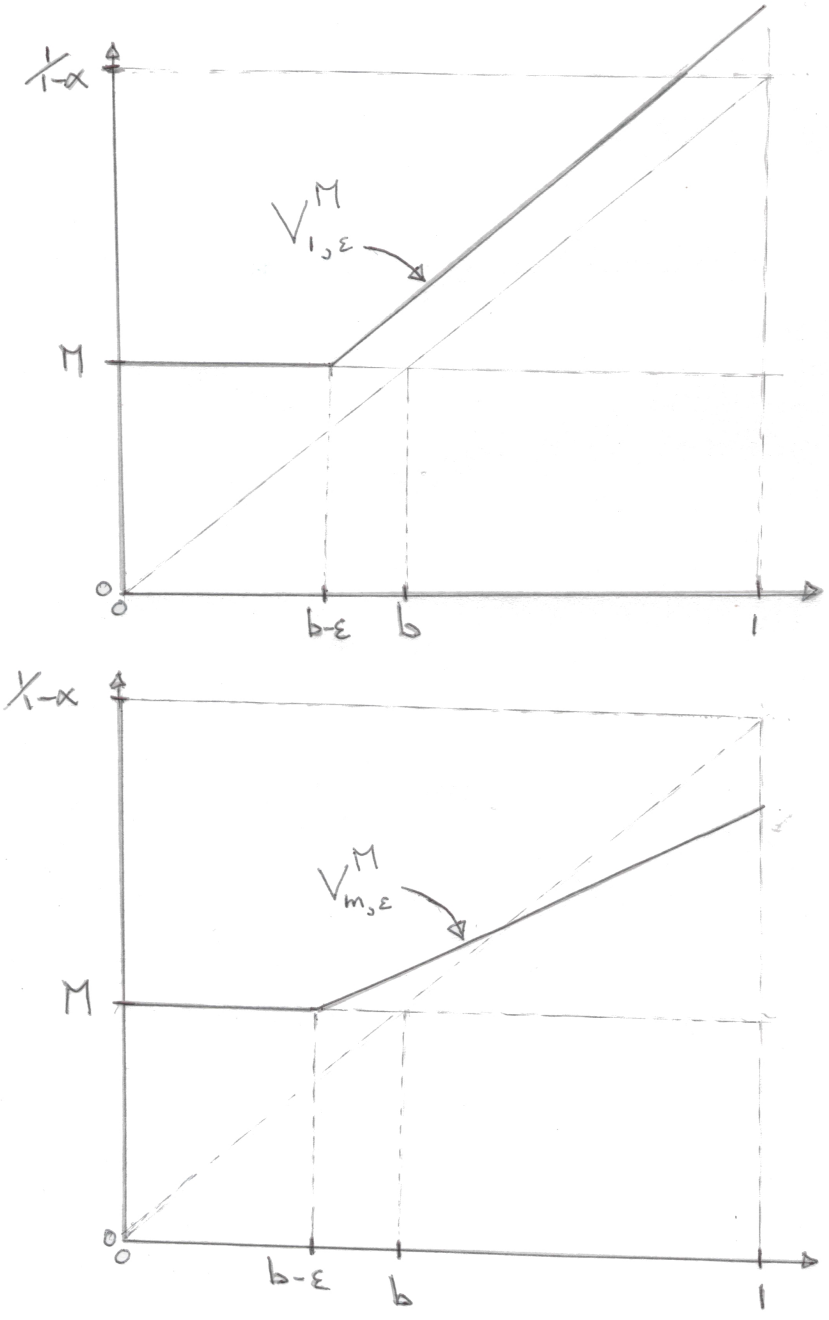


Figure 3.3: Examples of the family of piecewise linear functions of the form $V^M_{\epsilon,m}$.

and that $T_n\left(V^M_{m,\epsilon}(.)\right)$ is linear in $\left[\frac{1}{n-1}, \frac{n-2}{n-1}\right] \setminus I_{b,\epsilon,n}$. Moreover

$$T_n\left(V^M_{m,\epsilon}(x)\right) \geq V^M_{m,\epsilon}(x)$$

and $T_n\left(V^M_{m,\epsilon}(x)\right)$ is $\frac{m}{1-\alpha}$-Lipschitz.

Now we directly compute the effects of $L \circ T_n$ on the functions $V^M_{m,\epsilon}$:

Lemma 3.14 *For all $n > 2$ and $x \in I_{b,\epsilon,n}$,*

$$(L \circ T_n)\left(V^M_{m,\epsilon}(x)\right) = \frac{b\alpha}{1-\alpha} + \frac{m\alpha}{1-\alpha} x \left(\frac{n-1}{n} x - (b-\epsilon) + \frac{1}{n} + \frac{1-\alpha}{m\alpha}\right)$$

and

$$\frac{d}{dx}(L \circ T_n)\left(V^M_{m,\epsilon}(x)\right) = 1 + \frac{m\alpha}{1-\alpha}\left(2\frac{n-1}{n} x - (b-\epsilon) + \frac{1}{n}\right).$$

Proof. This is elementary algebra. ■

While T_n has the effect of lifting $V^M_{m,\epsilon}(\cdot)$, L mostly has the effect of squashing $T_n\left(V^M_{m,\epsilon}(.)\right)$ vertically. This observation leads us to find the first stability result which will be instrumental in bounding V^M above, and obtaining the Upper Bound. It isolates, for fixed values of M and m, the value of ϵ which remains stable under $(L \circ T_n)$

Corollary 3.15 *Fix $\frac{1}{n} < \frac{1-\alpha}{m\alpha}$. Then*

$$(L \circ T_n)\left(V^M_{m,\epsilon}(b-\epsilon)\right) > M$$

$$\begin{aligned} \iff \epsilon < \epsilon_{n,b,m} :&= \frac{\left(\frac{1-\alpha}{m\alpha} n - (2b-1)\right)}{2}\left[\left(1 + \frac{4b(1-b)}{\left(\frac{1-\alpha}{m\alpha} n - (2b-1)\right)^2}\right)^{\frac{1}{2}} - 1\right] \\ &= \frac{1}{n}\frac{m\alpha}{1-\alpha} b(1-b) + O\left(\frac{1}{n^2}\right), \end{aligned}$$

and the equality holds at $\epsilon_{n,b,m}$.

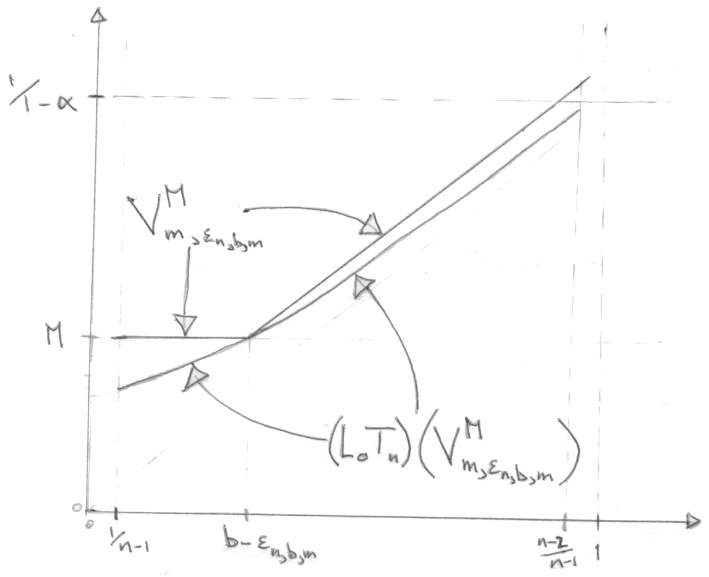


Figure 3.4: Illustration of Stability in Corollary 3.15

Proof. Taking $x = b - \epsilon$ in Lemma 3.14 we find that $(L \circ T_n)\left(V^M_{m,\epsilon}(b-\epsilon)\right) < M$ yields a quadratic inequality in ϵ, and the result follows from elementary computations. ■

The first part of the following Corollary tracks the movement of the discontinuity in the first derivative of $V^M_{m,\epsilon}$ under the operator $(W \circ L \circ T_n)$. The second part shows a stability result on the right-hand derivative at this point, again under the operator $(W \circ L \circ T_n)$. This stability property will be instrumental in the proof of the Lower Bound.

Given u, $v \in \mathbb{R}$, we use the notation $u = \square(v)$ iff $|u| \leq v$. Let

$$c := \frac{\alpha b(1-b)}{1-\alpha b}.$$

Fix $D > 0$, which will be chosen later, and let

$$m(D, b, \alpha, n) = \frac{1-\alpha}{1-b\alpha + D/n}.$$

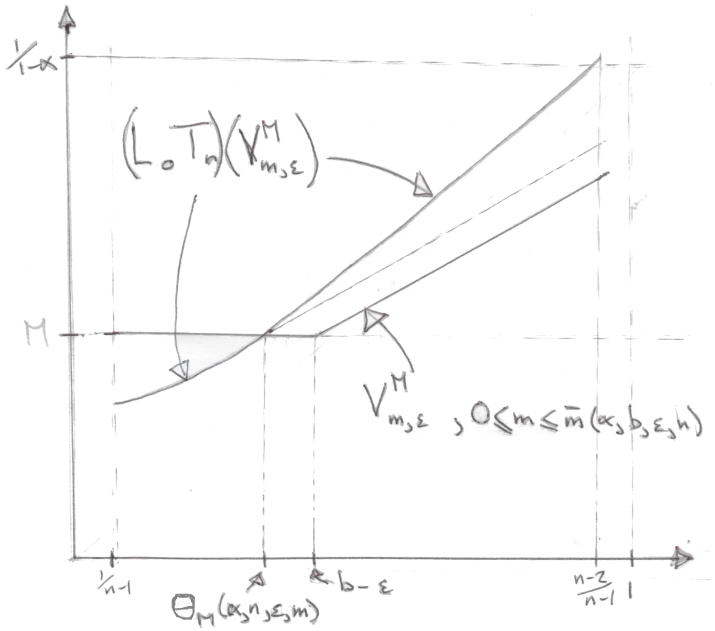


Figure 3.5: Illustration of Stability in Corollary 3.16

Corollary 3.16 *Fix $n \geq 2$, and $D > 0$. Assume $\epsilon \leq D/n$, let $m := m(D, b, \alpha, n)$, and let $\theta := \theta_M(\alpha, n, \epsilon, m) > 0$ be the smallest solution, in y, of*

$$(L \circ T_n)\left(V^M_{m,\epsilon}(b-y)\right) = M.$$

Then

$$(n-1)\theta - n\epsilon = -(\alpha b - 1)(c - n\epsilon) + \square\left(\frac{\mathsf{Cst}(D, b, \alpha)}{n}\right)$$

where $\mathsf{Cst}(D, b, \alpha)$ denotes a constant given values of D, b, and α. Moreover

$$\frac{d}{dx}\left[(L \circ T_n)\left(V^M_{m,\epsilon}(b-\theta)\right)\right] > \frac{m}{1-\alpha}.$$

Proof. It follows from Lemma 3.14 that θ is solution of the quadratic equation

$$y^2 - B_n y + C_n = 0, \tag{3.13}$$

where

$$B_n := 2b + \frac{n}{n-1}\left(\frac{1-\alpha}{\alpha m} + \frac{1}{n} - (b-\epsilon)\right) = \frac{1}{\alpha} + \square\left(\frac{\mathsf{Cst}(D,b,\alpha)}{n}\right)$$
$$C_n := \frac{b(1-b+n\epsilon)}{n-1}.$$

Then

$$\theta = \frac{B_n - \sqrt{B_n^2 - 4C_n}}{2} = \frac{B_n\left(1-\sqrt{1-4C_n/B_n^2}\right)}{2} = \frac{C_n}{B_n} + \square\left(\frac{\mathsf{Cst}(D,b,\alpha)}{n}\right).$$

The corollary follows by elementary calculations, using Lemma 3.14 in the second part. ■

3.3.3 The Upper Bound

Fix α and $n \geq 2$. Since the index is non-decreasing in x,

$$\left(x \leq x_{\max}(n, M, \alpha)\right) \implies \left(\nu\left(x, \frac{1}{n}\right) \leq M(1-\alpha)\right).$$

Thus in order to upper bound the index $\nu\left(x, \frac{1}{n}\right)$, we need to obtain a lower bound $x_{\max}(n, M, \alpha)$, which will be a consequence of an upper bound for V^M of the form $V^M_{m,\epsilon}$:

Proposition 3.17 *Fix α and $M = \frac{b}{1-\alpha}$, and let $\frac{1}{n} < \frac{1-\alpha}{\alpha}$. Then for all $\left(x, \frac{1}{n}\right) \in \mathbb{G}_n$*

$$V^M_{1,\epsilon_{n,b,1}}(x) \geq V^M\left(x, \frac{1}{n}\right)$$

where $\epsilon_{n,b,1}$ is as defined in Corollary 3.15.

Proof. Fix $M = \frac{b}{1-\alpha}$ and let $\frac{1}{n} < \frac{1-\alpha}{\alpha}$.

First we show, by descending induction on $n \leq n_0$, that

$$V^M_{1,\epsilon_{n,b,1}}(x) \geq V^{M,n_0}\left(x, \frac{1}{n}\right)$$

for all $\frac{1}{n_0} < \frac{1}{n} < \frac{1-\alpha}{\alpha}$. We then conclude the result as an application of Lemma 3.10.

Note that

$$V^{M,n_0}\left(x,\frac{1}{n_0}\right)\leq V^{M}_{1,\epsilon_{n_0,b,1}}(x)$$

Using the assumption that $\frac{1}{n_0}<\frac{1}{n}<\frac{1-\alpha}{\alpha}$, from Corollary 3.15 we know that

$$(L\circ T_{n+1})\left(V^{M}_{1,\epsilon_{n+1,b,1}}(b-\epsilon_{n+1,b,1})\right)=\frac{b}{1-\alpha}$$

and so since by Lemma 3.9, $(L\circ T_{n+1})\left(V^{M}_{1,\epsilon_{n+1,b,1}}(\cdot)\right)$ is $\frac{1}{1-\alpha}$-Lipschitz we have

$$(L\circ T_{n+1})\left(V^{M}_{1,\epsilon_{n+1,b,1}}(x)\right)\leq V^{M}_{m,\epsilon_{n,b,1}}(x)$$

Hence if $V^{M,n_0}\left(x,\frac{1}{n+1}\right)\leq V^{M}_{1,\epsilon_{n+1,b,1}}$, using that $(W\circ L\circ T_{n+1})$ is nonnegative, we obtain

$$\begin{aligned}V^{M,n_0}\left(x,\frac{1}{n}\right)&=(W\circ L\circ T_{n+1})\left(V^{M,n_0}\left(x,\frac{1}{n+1}\right)\right)\\&\leq(W\circ L\circ T_{n+1})\left(V^{M}_{1,\epsilon_{n+1,b,1}}(x)\right)\\&\leq V^{M}_{1,\epsilon_{n,b,1}}(x).\end{aligned}$$

This concludes the induction step, and Lemma 3.10 now gives the result when we take the limit as n_0 goes to infinity. ■

Proof of Theorem 3.2 part 1. Fix $\frac{1}{n}\leq\frac{1-\alpha}{\alpha}$. Proposition 3.17 implies $x_{\max}(n,M,\alpha)\geq b-\epsilon_{n,b,1}$ so that, by (3.9),

$$\nu\left(x,\frac{1}{n}\right)\leq\inf\{b:x\leq b-\epsilon_{n,b,1}\}.$$

Now an explicit computation on $\epsilon_{n,b,1}$ (defined in Corollary 3.15) shows that

$$x\leq b-\epsilon_{n,b,1}\implies b\geq x+\frac{\alpha}{1-\alpha}x(1-x)\frac{1}{n},$$

which enables us to conclude. ■

3.3.4 The Lower Bound

As in the case for the Upper Bound first we fix α and $n\geq 2$. Then

$$(x>x_{\max}(n,M,\alpha))\implies\left(\nu\left(x,\frac{1}{n}\right)>M\right).$$

Similarly as in the upper bound of the index, we prove a lower bound for V^M of the form $V^M_{m,\epsilon}$, in order to upper bound $x_{\max}(n, M, \alpha)$.

Given α, b, $D > 0$ and n_0, let us define the following sequence $(\epsilon_n^{n_0})_{n\le n_0} := (\epsilon_n^{n_0}(D, b, \alpha, n))_{n\le n_0}$ by descending induction: $\epsilon_{n_0}^{n_0} := 0$ and, for all $n \le n_0$,

$$\epsilon_{n-1}^{n_0} := \theta_M(\alpha, n, \epsilon_n^{n_0}, m(D, b, \alpha, n)).$$

Proposition 3.18 *Fix α, $M \in \left[0, \frac{1}{1-\alpha}\right]$, and $D > 2c$. Let $\epsilon_n^{n_0} := \epsilon_n^{n_0}(D, b, \alpha, n)$, $m := m(D, b, \alpha, n)$. For all $\left(x, \frac{1}{n}\right) \in \mathbb{G}_n$ with n_0, $n \ge \mathsf{Cst}(D, b, \alpha)$ then*

$$V^{M,n_0}\left(x, \frac{1}{n}\right) \ge V^M_{m,\epsilon_n^{n_0}}(x),$$

and

$$\epsilon_n^{n_0} \le \frac{D}{n}, \ \limsup_{n_0\to\infty} \epsilon_n^{n_0} \le \frac{c}{n} + \frac{\mathsf{Cst}(D, \alpha, b)}{n^2}.$$

Proof. Let us first prove the second part of the proposition. Fix n_0, let $\lambda := \alpha b < 1$ and let, for all $n \le n_0$, $U_n := c - n\epsilon_n^{n_0}$. By Corollary 3.16, for all $n \le n_0$, if $\epsilon_n^{n_0} \le D/n$,

$$U_{n-1} = \lambda U_n + F_n/n,$$

where $|F_n| \le \mathsf{Cst}(D, b, \alpha)$.

Therefore, by descending induction, for all $n \le n_0$, as long as $\epsilon_n^{n_0} \le D/n$,

$$U_n = c\lambda^{n_0-n} + \sum_{j=n+1}^{n_0} \frac{F_j\lambda^{j-n-1}}{j}.$$

Now, for all $n \ge 1$,

$$\sum_{j=n+1}^{n_0} \frac{\lambda^{j-n-1}}{j} \le \frac{1}{n+1} + \int_0^\infty \frac{\lambda^x}{x+n}\,dx \le \frac{\mathsf{Cst}(\lambda)}{n},$$

which enables us to conclude the second part if $n_0 \ge n \ge \mathsf{Cst}(D, b, \alpha)$.

The proof of the first part of the proposition is, as in the proof of Proposition 3.17, based on a descending induction, which is here not only on the zeros of $V^{M,n_0}(1/n, .) - M$, but also on the derivatives.

Let $m := m(D, b, \alpha, n)$ and $\epsilon_n^{n_0} := \epsilon_n^{n_0}(D, b, \alpha, n, n_0)$. We prove that, for all $n \leq n_0$, $1/n \leq x < y \leq (n-1)/n$,

$$V^{M,n_0}\left(x, \frac{1}{n}\right) \geq V^M_{m,\epsilon_n^{n_0}}(x) \tag{3.14}$$
$$\Delta_{x,y}\left[V^{M,n_0}\left(\cdot, \frac{1}{n}\right)\right] \geq \Delta_{x,y}\left[V^M_{m,\epsilon_n^{n_0}}(\cdot)\right]$$

The induction step follows from the remark that, if (3.14) holds, then

$$\Delta_{x,y}\left[(L \circ T_n)\left(V^{M,n_0}\left(\cdot, \frac{1}{n}\right)\right)\right] \geq \Delta_{x,y}\left[(L \circ T_n)\left(V^M_{m,\epsilon_n^{n_0}}(\cdot)\right)\right],$$

and from Corollary 3.16. ■

Proof of Theorem 3.2 part 2. As in the proof of part 1, we deduce from (3.10) that

$$\nu_\alpha\left(x, \frac{1}{n}\right)) > (1-\alpha)\sup\left\{M : x_{\max}(n, M, \alpha) > b - \limsup_{n_0 \to \infty} \epsilon_n^{n_0}\right\},$$

and $\limsup_{n_0\to\infty} \epsilon_n^{n_0} = \alpha b(1-b)/(1-\alpha b) + O(1/n^2)$ by Proposition 3.18, so that the conclusion follows similarly. ■

3.4 A Preliminary Non-Zero Inconsistency Bound

We now prove Proposition 3.3 as a simple application of the upper bound, ν_U, for the Gittins Index developed in the previous section.

Assume throughout this section that A2 holds, that $p_1 > p_2$, $A = \{1, 2\}$, and $I_0 = \{(0,0), (0,0)\}$. Let E denote the event

$$E := \{A_n = 1 \text{ finitely often}\}.$$

In Section 3.2.2 we showed that $\mathbb{P}(E) > 0$. We shall show below a lower bound on the rate at which $\mathbb{P}(E)$ tends to zero as $\alpha \to 1$ when $p_2 \leq \frac{1}{2}$.

By analogy with Section 3.2.2, we define, for each $a \in A$ and $n \geq 0$ the $\mathcal{F}_n$-measurable random variable

$$\bar{x}_{a,n} := \frac{1 + \sum_{k=1}^n O_{a,k}}{2+n}.$$

Let $p_2 > \epsilon > 0$ and denote by $E_{1,\epsilon}$ and $E_{2,\epsilon}$ the events

$$E_{1,\epsilon} := \left\{ \nu_U \left(\bar{x}_{1,n}, \frac{1}{n} \right) \leq p_2 - \epsilon \text{ for some } n \geq \frac{\alpha}{1-\alpha} \right\}$$

$$\text{and } E_{2,\epsilon} := \left\{ \nu \left(\bar{x}_{2,n}, \frac{1}{n} \right) > p_2 - \epsilon \text{ for all } n \geq 1 \right\}.$$

Now, since under assumption A2 the sequences $\left\{ (O_{a,n})_{n \geq 1} \right\}_{a=1,2}$ are both i.i.d. and independent of each other, it follows that

$$\mathbb{P}(E) > \mathbb{P}(E_{1,\epsilon} \cap E_{2,\epsilon}) = \mathbb{P}(E_{1,\epsilon}) \mathbb{P}(E_{2,\epsilon}). \tag{3.15}$$

We can bound $\mathbb{P}(E_{1,\epsilon})$ below using the upper bound ν_U: Define, for each $C \in (0,1)$, the set

$$S_C := \{(x,t) \in \mathbb{G} : \nu_U(x,t) \leq C\}.$$

Defining also

$$N_C := \max \left\{ t^{-1} : (t,t) \in S_C \right\} > 0$$

a crude lower bound of $\mathbb{P}(E_{1,\epsilon})$ is given by

$$\begin{aligned} \mathbb{P}(E_{1,\epsilon}) &\geq \mathbb{P}\left(\bar{x}_{1,N_{p_2-\epsilon}} = N_{p_2-\epsilon} \right) \\ &= (1-p_1)^{N_{p_2-\epsilon}} \end{aligned}$$

when $N_{p_2-\epsilon} > \frac{\alpha}{1-\alpha}$ (recall that the upper bound is valid only when $t \leq \frac{1-\alpha}{\alpha}$). As a trivial consequence, if $N_{p_2-\epsilon} \leq \frac{\alpha}{1-\alpha}$ then

$$\mathbb{P}(E_{1,\epsilon}) \geq (1-p_1)^{\frac{\alpha}{1-\alpha}}.$$

In the Lemma below we upper bound $N_{p_2-\epsilon}$ and hence lower bound $\mathbb{P}(E_{1,\epsilon})$. However in order for our argument in this Lemma to hold we require the extra assumption that $p_2 - \epsilon \leq \frac{1}{2}$:

Lemma 3.19 *When $p_2 - \epsilon \leq \frac{1}{2}$,*

$$\mathbb{P}(E_{1,\epsilon}) \geq (1-p_1)^{f(p_2,\epsilon,\alpha)}$$

where

$$f(p_2,\epsilon,\alpha) := \max \left\{ \left\lceil \frac{(p_2-\epsilon)(1-(p_2-\epsilon)) + (1-\alpha)}{(1-\alpha)(p_2-\epsilon)} \right\rceil, \frac{\alpha}{(1-\alpha)} \right\}.$$

Proof. Fix $\alpha \in (0,1)$ and $C \in \left(0, \frac{1}{2}\right)$.

To find an upper bound for N_C we rearrange the curve $C = \nu_U(x,t)$ to obtain

$$t(x) = \frac{(C-x)(1-\alpha)}{x(1-x)}$$

Then

$$\frac{d}{dx}t(x) = -(1-\alpha)\left[\frac{C - x(2C-x)}{[x(1-x)]^2}\right]$$

and

$$\frac{d^2}{dx^2}t(x) = 2(1-\alpha)\frac{(C(1-3x(1-x)) - x^3)}{[x(1-x)]^3}.$$

So when $x \leq C \leq \frac{1}{2}$

$$\frac{d}{dx}t(x) \leq 0 \text{ and } \frac{d^2}{dx^2}t(x) \geq 0.$$

Hence the tangent to $t(C,\alpha,x)$ at $x = C$, which we shall denote $\bar{t}(C,\alpha,x)$, satisfies

$$\bar{t}(x) \leq t(x) \text{ for all } x \leq C$$

Now

$$\bar{t}(x) = \left[\frac{d}{dx}t(C)\right](x-C) = -\frac{(1-\alpha)}{C(1-C)}(x-C).$$

So $\bar{t}(x) = x$ iff

$$\left[1 + \frac{1-\alpha}{C(1-C)}\right]x = \frac{1-\alpha}{1-C}$$
$$\iff x = \frac{(1-\alpha)C}{C(1-C)+(1-\alpha)}.$$

Hence it follows from the definition of N_C that

$$N_C \geq \left\lceil \frac{C(1-C)+(1-\alpha)}{(1-\alpha)C} \right\rceil.$$

Now the upper bound $\nu_U(x,t)$ is only valid when $t < \frac{1-\alpha}{\alpha}$. Hence we may only conclude that $\mathbb{P}(E_{1,\epsilon}) \geq (1-p_1)^{f(p_2,\epsilon,\alpha)}$ where $f(p_2,\epsilon,\alpha)$ is as defined in the statement of the Lemma. ■

Now we turn our attention to $\mathbb{P}(E_{2,\epsilon})$. Since the lower bound ν_L developed in the previous section is only valid for large enough n, we work instead with the lower bound given in Lemma 3.1

$$\nu\left(x, \frac{1}{n}\right) \geq x$$

which holds for all $\left(x, \frac{1}{n}\right) \in \mathbb{G}$. We therefore consider the event

$$E'_{2,\epsilon} := \{\bar{x}_{2,n} > p_2 - \epsilon \text{ for all } n \geq 1\} \, (\subset E_{2,\epsilon}) \, .$$

By a similar application of the Strong Law of Large Numbers to that found in the proof of Theorem 3.7 (specifically the proof that $\mathbb{P}(F_{\epsilon,t}(n_0)) > 0$) to show that

$$\left(p_2 - \epsilon \leq \frac{1}{2}\right) \implies \left(\mathbb{P}\left(E'_{2,\epsilon}\right) > 0\right). \tag{3.16}$$

(The argument proceeds by applying the Strong Law of Large Numbers to show that, with positive probability, after some time n_0, $x_{2,n} > p_2 - \epsilon$ for all $n \geq n_0$. It then concludes by showing that on a non-null subset of this event $x_{2,n} > p_2 - \epsilon$ also holds for all $n < n_0$.)

We thus deduce the lower bound on the rate at which $\mathbb{P}(E) \to 0$ as $\alpha \to 1$ given in Proposition 3.3:

Proof of Proposition 3.3. A calculation shows that, for fixed $C > 0$

$$\frac{C(1-C) + (1-\alpha)}{(1-\alpha)C} > \frac{\alpha}{1-\alpha}$$
$$\iff C < \frac{1}{2}\left[(1-\alpha) + \sqrt{(1-\alpha)(5-\alpha)}\right].$$

So $\lim_{\alpha \to 1} \left[f(p_2, \epsilon, \alpha) / (\alpha/1-\alpha)\right] = 1$, and the result follows from (3.15), Lemma 3.19 and (3.16). ■

Remark 3.20 *There is much room for improvement in this result, and it is the subject of current research. Instead of Lemma 3.19 we could take the upper bound for $N_{p_2-\epsilon}$ given by the solution lying between 0 and 1 of the equation*

$$x = \frac{(C-x)(1-\alpha)}{x(1-x)}.$$

However this is cubic in x and its solutions do not present a convenient algebraic form to work with. Similarly it is possible to obtain explicit lower bounds for $\mathbb{P}(E_{2,\epsilon})$, but at present these are only given as infinite products.

3.5 Conclusion

This work was motivated through attempting to understand how questions of learning relate to questions of reward maximisation in Bandit Problems. In particular we have studied the Gittins Index Strategy which maximises total expected reward for Bandits with bounded, i.i.d., and geometrically discounted rewards. In the case of Bernoulli reward sequences we have achieved new asymptotic upper and lower bounds for the Gittins Indices which come close to giving the second term in the Taylor expansion of the indices around time $n = \infty$. Exactly what this second term is remains an open question.

In addition for the case when we have just two actions and when the success probability of the worse action is smaller than $\frac{1}{2}$, we have also achieved a provisionary lower bound for the rate of decay of the inconsistency of the Gittins Index Strategy as $\alpha \to 1$ through an application of the new upper bound, although there is much room for improvement here. In particular it is not clear from our analysis how this quantity is affected by the size of the difference $p_1 - p_2$.

Bibliography

[1] J. C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in Satistics*, pages 241 – 266, 1974a.

[2] John Gittins. *Multi-Armed Bandit Allocation Indices*. Wiley, 1989.

[3] N. El Karoui and I. Karatzas. Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, 2:255 – 286, 1994.

[4] H. Kaspi and A. Mandelbaum. Multi-armed bandits in descrete and continuous time. *The Annals of Applied Probability*, 8:1270–1290, 1998.

[5] F. Kelly. Multi-armed bandits with discount factor near one: The bernoulli case. *The Annals of Statistics*, 9:987–1001, 1981.

[6] P. Whittle. *Optimization Over Time*, volume 1. Wiley, 1982.

Chapter 4

A Consistency and Regret Analysis of the Thompson Strategy for Bandit Problems

Abstract

We give an elementary proof of the consistency of the Thompson Strategy for N-Armed Bandits under a mild model consistengy condition, and we provide a scheme for obtaining both cumulative and simple regret bounds for this strategy. In particular we find that for 2-Armed Bandits with Bernoulli payoffs the Thompson Strategy achieves logarithmic cumulative regret.

4.1 Introduction

In a Bandit problem an agent is presented with a number of possible actions each of which produce repeatable, but unknown (random) behaviour upon repeated implementation. Hence through repeated application of any particular action the agent will gain information about the behaviour that action produces. However at any time the agent must choose to implement just one of these actions and therefore faces a dilemma: either he can choose to implement an action so as to maximise the chance, given his previous observations, of producing desirable behaviour, or he can choose to experiment by implementing an action that is more likely to produce undesirable behaviour. This is often referred to as an exploration-exploitation dilemma.

In this chapter we study the first strategy proposed for a Bandit problem, the Thompson Algorithm [6]. In particular we give a general proof of consistency and engage in a deeper analysis of the strength of this consistency. It is surprising that despite being proposed almost 80 years ago this Bayesian approach has not been treated to such limiting analyses in the literature. However it is often difficult to provide asymptotic analyses of Bayesian algorithms due, in part, to difficulties in finding simple closed form expressions for the integrals involved.

After showing consistency of the algorithm under a mild model consistency assumption, we demonstrate a method for analysing the asymptotic growth of the amount of time spent implementing a sub-optimal action. This allows us to bound both the rate of growth of the cumulative regret (a measure of long term value of a strategy first introduced in [4]) and the rate of decay of its simple regret (a measure of exploration efficiency recently introduced in [1]).

Model Let us suppose that we have a finite set of possible *actions* $A := \{1, \ldots, N\}$. We attach to each action, a, an i.i.d. sequence of $\mathcal{O}$-valued random variables $(O_{a,n})_{n\geq 1}$, where $\mathcal{O} \subseteq \mathbb{R}$ is some *observation space*. After choosing action a the agent observes $O_{a,n}$, which we interpret as his *reward* at time n. As the agent makes his choice at each time he creates a sequence of actions and observations

$$I_0, (A_1, I_1), (A_1, I_1), \ldots, (A_n, I_n), \ldots$$

where I_0 represents the initial information known about the actions, A_n is the action taken at time n, and $I_n := O_{A_n,n}$ is the observation made, or reward gained at time n.

Let

$$\mathcal{X}_n := [I_0, (A_1, I_1), (A_1, I_1), \ldots, (A_n, I_n)]$$

and let $\mathcal{H}$ be the set of all possible partial histories of the form $\mathcal{X}_n$. Then a *strategy*, π, is defined to be a map $\pi : \mathcal{H} \to \mathcal{P}(A)$, where $\mathcal{P}(A)$ denotes the set of probability distributions over A. Thus given the agent's choice of strategy and initial information we can view $(\mathcal{X}_n)_{n\geq 0}$ as a random process evolving in $\mathcal{H}$ with transition probabilities:

$$\mathbb{P}((A_{n+1}, I_{n+1}) = (a, o) \in A \times \mathcal{O} \mid \mathcal{F}_n) = \pi(\mathcal{X}_n)(a) . \mathbb{P}(O_{a,n+1} = o)$$

where

$$\mathcal{F}_n = \sigma\left(I_0, (A_k)_{n\geq k\geq 1}, \left\{(O_{a,k})_{n\geq k\geq 1}\right\}_{a\in A}\right).$$

We call $(\mathcal{X}_n)_{n\geq 0}$ the *Bandit process*.

The Thompson Strategy We assume that for each a the common distribution of the i.i.d. sequence $(O_{a,n})_{n\geq 1}$comes from some known parametrised family of distributions with parameters in some one-dimensional parameter space Θ. For example we might assume that each action produces a sequence of Bernoulli trials with a different, fixed, but unknown success probability. Or we might assume that each action, a, produces uniform random numbers on some interval $[0, x_a]$ where the parameters $\{x_a\}_{a\in A}$ are unknown. In this context, given the initial information I_0, we can choose prior distributions for the parameters of the distribution of $O_{a,1}$, and compute posterior distributions for these parameters given the observations $\mathcal{X}_n$.

Hence, for each a, we define $(G_{a,n})_{n\geq 0}$ to be the posterior distribution for the unknown quantity $p_a := \mathbb{E}(O_{a,1})$, given the initial information I_0, a choice of priors, and the observations $\mathcal{X}_n$. Let $\hat{p}_{a,n}$ denote a random variable with distribution $G_{a,n}$ and write

$$\bar{p}_{a,n} := \mathbb{E}(\hat{p}_{a,n} \mid \mathcal{X}_n)$$

Then the Thompson strategy, π_{Th}, is defined by

$$\begin{aligned}\pi_{Th}(\mathcal{X}_n)(a) &:= \mathbb{P}\left(\hat{p}_{a,n} = \max_{k\in A} \hat{p}_{k,n} \mid \mathcal{X}_n\right) \\ &= \int_{u_1,\ldots,u_N} \mathbf{1}_{\{u_a = \max_{k\in A} u_k\}} \prod_{i\in A} dG_{i,n}(u_i). \end{aligned} \tag{4.1}$$

so that the probability of choosing to implement action a at time n is the posterior probability that action a yields the greatest expected reward.

Consistency A strategy is said to be *consistent* if it satisfies the following property:

$$\mathbb{P}_\pi\left(\lim_{n\to\infty}\left(\frac{\sum_{k=1}^n I_k}{n}\right) = \max_{a\in A}\{p_a\}\right) = 1.$$

Thus a strategy is consistent iff its average rewards are as good as the optimal reward in the limit. Otherwise it is said to be *inconsistent*.

Consistency is a classic learning condition required of a Bandit Problem strategy, however it is not a good measure of learning optimality as there are many strategies which are consistent. One measure by which to distinguish between consistent strategies is that of *regret*.

Regret The regret of a strategy at time n, first introduced and studied by Lai and Robbins in [4], is defined to be

$$R_\pi(n) := \mathbb{E}\left(\sum_{k=1}^n O_{a^*,n} - I_n\right),$$

where $a^* \in A_{\max} := \arg\max_{a\in A}\{p_a\}$. This definition of regret is sometimes called the *cumulative regret*, and it was shown in [4] that for any strategy satisfying

$$R_\pi(n) = o(n^\alpha) \text{ for all } \alpha \in (0,1), \text{ almost surely,}$$

the cumulative regret must growing at

$$\liminf_{n\to\infty} \frac{\mathcal{R}_\pi(n)}{\log n} \geq \sum_{p_j \neq p_{\max}} \frac{p_{\max} - p_j}{D\left(p_j, p_{\max}\right)},$$

where $D\left(p_j, p_{\max}\right)$ is the Kullback-Leibler divergence between p_j and $p_{\max}$.

Results As a corollary of Theorems 4.6 and 4.12 we are able to show that when $A = \{1, 2\}$, and

$$\left(O_{1,a} \sim Bern\left(p_a\right) \text{ for } a = 1, 2\right) \text{ or } \left(O_{1,a} \sim Unif\left(0, x_a\right) \text{ for } a = 1, 2\right)$$

where the parameters $\{p_a\}_{a\in A}$ or $\{x_a\}_{a\in A}$ are unknown and no initial information is given,

- $R_{\pi_{Th}}(n) = O(\ln n)$ almost surely.

The chapter is organised as follows. In the next section we give the proof of consistency. In Section 4.3 we first describe our general approach to establishing regret bounds and then proceed to apply the method first to the case of uniform rewards and subsequently to the case of Bernoulli rewards. Finally we include an appendix with some technical results used in the chapter.

4.2 Consistency

Consider the following model consistency condition:

(C) Let $(U_n)_{n\geq 1}$ be any a sequence of i.i.d. random variables with common distribution in a set of single parameter distributions whose parameters lie in the parameter space Θ; let θ be the parameter of the common distribution of $(U_n)_{n\geq 1}$, and let $Y_n := (U_1, \ldots, U_n)$. We shall say that a prior distribution ρ_0 for θ satisfies (C) if $supp\left(\rho_0\right) = \Theta$ and the posterior distributions ρ_n derived from ρ_0 and the observations Y_n satisfy

$$\rho_n \to \delta_\theta \text{ in } L^1 \text{ as } n \to \infty \text{ almost surely.}$$

We shall assume throughout the chapter that the strategy chooses prior distributions so that

A1 $G_{a,0}$ satisfies (C) for all $a \in A$.

This assumption is often easy to satisfy. For example, in the case when we assume that our actions yield Bernoulli rewards it is sufficient to take $G_{a,0}$ as a uniform distribution on $[0,1]$, for each $a \in A$. Also when the actions yield uniform rewards on intervals $[0, x_a]$ for unknown values of x_a we can take non-informative priors satisfying A1.

Assuming A1 holds, if under π_{Th} the agent implements all actions infinitely often, then clearly the strategy will be consistent.

We therefore show the following:

Lemma 4.1 *Assuming A1 holds, under the strategy π_{Th} all actions are implemented infinitely often almost surely.*

Proof. We may assume, without loss of generality, that for some $k \in A$ we implement actions $1, \ldots, k-1$ only finitely often, but implement actions $k, \ldots, N$ infinitely often.

Let E be the event on which $k > 1$. Then for each $k \leq a \leq N$

$$G_{a,n} \to \delta_{p_a} \text{ as } n \to \infty \text{ almost surely}$$

almost surely on E. However since on E for each $1 \leq a < k$ we implement action a only finitely often,

$$G_{a,n} = G_{a,n_a} \text{ for all } n \geq n_a$$

for some finite, random time n_a.

So on E

$$\begin{aligned}
&\pi_{Th}\left(\mathcal{X}_n\right)\left(\{1,\ldots,k-1\}\right) \\
&= \int_{u_1,\ldots,u_N} \sum_{a\in\{1,\ldots,k-1\}} \mathbf{1}_{\{u_a=\max_{i\in A} u_i\}} \prod_{j\in A} dG_{j,n}\left(u_j\right) \\
&\to \int_{u_1,\ldots,u_N} \sum_{a\in\{1,\ldots,k-1\}} \mathbf{1}_{\{u_a=\max_{i\in A} u_i\}} \prod_{j=1}^{k-1} dG_{j,n_j}\left(u_j\right) \prod_{j=k}^{n} d\delta_{p_j}\left(u_j\right),
\end{aligned}$$

almost surely, and so in particular, using the assumption that $support(G_{a,1}) = \Theta$, for all $n \geq \max_{i \in \{1...,k-1\}} n_i$

$$\pi_{Th}\left(\mathcal{X}_n\right)\left(\{1,\ldots,k-1\}\right) \geq \epsilon$$

where ϵ is some random, positive constant. It follows from the Conditional Borel-Cantelli Lemma (see [7], pg. 124) that $A_n \in \{1,\ldots,k-1\}$ infinitely often on E, almost surely, which contradicts the definition of E almost everywhere on E.

Hence E is null and we find that π_{Th} implements each action infinitely often almost surely. ■

As discussed above the result now follows immediately:

Theorem 4.2 *Under the assumption A1 the Thompson strategy is consistent.*

4.3 Regret Bounds

4.3.1 General Method

We make the simplifying assumptions that $A = \{1,2\}$, that the agent is given no initial information, and that $p_1 > p_2$. We also make the notation

$$t_{a,n} := \sum_{1 \leq k \leq n} \mathbf{1}_{\{A_k = a\}}.$$

In this setting the cumulative regret of a strategy can be written simply as

$$R_\pi\left(n\right) = \left(p_1 - p_2\right)\mathbb{E}_\pi\left(t_{2,n}\right)$$

Now writing

$$X_n := \pi_{Th}\left(\mathcal{X}_n\right)\left(1\right),$$

and

$$\xi\left(n\right) := \sum_{0 \leq k \leq n-1} \left(1 - X_k\right)$$

we note the following lemma:

Lemma 4.3 *Under assumption A1*

$$\lim_{n\to\infty} \frac{t_{2,n}}{\xi(n-1)} = 1$$

almost surely.

Proof. By Lemma 4.1 under assumption A1 the strategy implements action 2 infinitely often, almost surely. Hence by the Conditional Borel Cantelli Lemma (see for example [7], page 124) $\xi(n) \to \infty$ as $n \to \infty$. So the result follows by the Strong Law of Large Numbers for martingales (see for example [7], page 123). ■

So we can make a study of both the cumulative and simple regrets through studying the growth of $\xi(n)$.

Let

$$\Delta\xi(n) := \xi(n+1) - \xi(n) = (1 - X_n).$$

Using (4.1) it is often possible to achieve an estimate of the form

$$\Delta\xi(n) \leq F(\xi(n))$$

for some function F which is positive, convex, and decreasing on the interval $(0, \infty)$. Thus it follows from the following proposition that $\xi(n) = O(f(n))$ where f is the solution of the initial value problem $f'(n) = F(f(n))$, $f(0) = \xi(0)$.

Proposition 4.4 *Suppose that the non-decreasing sequence* $(\xi(n))_{n\geq 1}$ *has* $\xi(0) \in (0,1)$ *and satisfies, for all large enough* n,

$$\xi(n+1) - \xi(n) \leq F(\xi(n)). \tag{4.2}$$

Suppose further that F *is convex, positive and decreasing on* $(0,\infty)$ *and consider the initial value problem*

$$f'(x) = F(f(x)),\ f(0) = \xi(0). \tag{4.3}$$

Then the solution, f*, of (4.3) exists uniquely and satisfies*

$$\xi(n) = O(f(n)).$$

We give the proof of this result in the Appendix.

Since estimating the value of $\Delta\xi(n)$ involves approximating the integral (4.1) we need to know the form of the posterior distributions $\{G_{a,n}\}_{a\in A,n\geq 1}$ to apply this scheme. In the following we focus on two cases; those of Uniform and Bernoulli rewards. However our method seems generally applicable when the distributions of the random variables $\{O_{1,a}\}_{a\in A}$ allow us to take conjugate prior families for the parameters of those distributions.

When it is defined it will be useful to write $g_{a,n}$ for the probability distribution function of $G_{a,n}$.

We will also frequently use the function

$$F_\delta : x \mapsto x - x^{\frac{1}{2}+\delta}.$$

and the following lemma:

Lemma 4.5 *Assume A1 holds. Then for any $\epsilon, \delta > 0$, for sufficiently large n*

$$(1-\epsilon) \leq \frac{t_{2,n}}{F_\delta(\xi(n))},$$

almost surely.

Proof. Note that

$$\sum_{k=1}^{n} Var\left(\mathbf{1}_{\{A_k=2\}} - X_{k-1}\right) = \sum_{k=1}^{n} X_{k-1}(1-X_{k-1}).$$

But by Lemma 4.1 under assumption A1 the strategy implements action 2 infinitely often, almost surely. Hence by the Conditional Borel Cantelli Lemma (see for example [7], page 124) $\xi(n) \to \infty$ as $n \to \infty$. Moreover, since under A1 π_{Th} is consistent, $X_n \to 1$ as $n \to \infty$. Hence

$$\sum_{k=1}^{\infty} Var\left(\mathbf{1}_{\{A_k=2\}} - X_{k-1}\right) = \infty,$$

and the result follows from the Law of the Iterated Logarithm for martingales (see for example [2]). ■

4.3.2 Uniform Rewards

We assume here that for each $a \in A$

$$O_{a,1} \sim Unif\,(0, x_a)$$

where the parameters $\{x_a\}_{a\in A}$ are unknown but are assumed to lie in $(0,\infty)$.

Let

$$y_{a,n} := \max_{1\le k\le n} \left\{ I_k \mathbf{1}_{\{A_k=a\}} \right\}.$$

Then taking the non-informative prior for $p_a = \frac{x_a}{2}$ we obtain the sequence of posterior distributions $\left(G_{a,n} \sim Pareto\left(t_{a,n}, \frac{y_{a,n}}{2}\right)\right)_{n\ge 1}$ where $G_{a,n}$ has the probability distribution

$$g_{a,n}(2x) = \begin{cases} 0, & 0 \le x \le y_{a,n}; \\ t_{a,n}\frac{y_{a,n}^{t_{a,n}}}{x^{t_{a,n}+1}}, & x > y_{a,n}. \end{cases}$$

Now in this case we can directly compute the value of X_n in terms of the random variables $y_{a,n}$ and $t_{a,n}$, $a = 1, 2$:

$$\begin{aligned} X_n &= \mathbb{P}\left(\hat{p}_{2,n} > \hat{p}_{1,n} \mid \mathcal{X}_n\right) \\ &= \mathbb{P}\left(2\hat{p}_{2,n} > 2\hat{p}_{1,n} \mid \mathcal{X}_n\right) \\ &= \int_{y=y_{1,n}}^{\infty} \left(\int_{x=y_{2,n}}^{y} t_{2,n} \frac{y_{2,n}^{t_{2,n}}}{x^{t_{2,n}+1}} dx \right) t_{1,n} \frac{y_{1,n}^{t_{1,n}}}{y^{t_n+1}} dy \\ &= \int_{y=y_{1,n}}^{\infty} \left(1 - \frac{y_{2,n}^{t_{2,n}}}{y^{t_{2,n}}} \right) t_{1,n} \frac{y_{1,n}^{t_{1,n}}}{y^{t_{1,n}+1}} dy \\ &= 1 - \int_{y=y_{1,n}}^{\infty} t_{1,n} \frac{y_{1,n}^{t_{1,n}} y_{2,n}^{t_{2,n}}}{y^{t_{1,n}+t_{2,n}+1}} dy \\ &= 1 - \frac{t_{1,n}}{t_{1,n}+t_{2,n}} \left(\frac{y_{2,n}}{y_{1,n}} \right)^{t_{2,n}}. \end{aligned}$$

So

$$1 - X_n = \frac{t_{1,n}}{t_{1,n}+t_{2,n}} \left(\frac{y_{2,n}}{y_{1,n}} \right)^{t_{2,n}} \tag{4.4}$$

Now since π_{Th} is consistent and, by Lemma 4.1, both actions are implemented infinitely often, almost surely, we have

$$\frac{t_{1,n}}{t_{1,n}+t_{2,n}} = \frac{t_{1,n}}{n} \to 1 \text{ and } \frac{y_{2,n}}{y_{1,n}} \to \frac{x_2}{x_1} < 1 \text{ a.s.}$$

as $n \to \infty$. Hence $(1 - X_n)$ is decaying exponentially in $t_{2,n}$.

It now follows from Lemma 4.5 that for all $0 < \epsilon < 1 - \frac{x_2}{x_1}$ and $\delta > 0$, for large enough n

$$\Delta\xi\left(n\right) \leq \exp\left(\ln\left(\frac{x_2}{x_1} + \epsilon\right) F_\delta\left(\xi\left(n+1\right)\right)\right) \text{ a.s..} \tag{4.5}$$

We can therefore conclude from Proposition 4.4 the following result:

Theorem 4.6 $\xi(n) = O(\ln n)$

Proof. Let $e > A > 1$.

It follows from (4.5) that for some small enough $\gamma > 0$, for all sufficiently large n,

$$\Delta\xi\left(n\right) \leq \frac{A}{\gamma} e^{-\gamma\xi(0)} \text{ a.s..}$$

Now the solution to the initial value problem

$$f(0) = \ln A,\ f'(x) = \frac{A}{\gamma} e^{-\gamma f(x)}$$

is $f(x) = \frac{1}{\gamma}\ln(x+1) + \ln A$. We thus conclude the result from Proposition 4.4. ■

4.3.3 Bernoulli Rewards

We assume here that for each $a \in A$

$$O_{a,1} \sim Bernoulli\left(p_a\right), p_a \in [0,1]$$

so that $O_{a,1} \in \{0,1\}$, and we define the random variables:

$$s_{a,n} := \sum_{1 \leq k \leq n} \mathbf{1}_{\{A_k = a\}} O_{a,k},\ \text{and}\ f_{a,n} := t_{a,n} - s_{a,n}.$$

We take $G_{a,n} \sim Beta\left(s_{a,n}+1, f_{a,n}+1\right)$, which is equivalent to taking the uniform prior for the parameter p_a, when no initial information is given. For this choice of prior

$$\bar{p}_{a,n} = \frac{s_{a,n}+1}{t_{a,n}+2}.$$

First we develop a simple and direct bound on ξ from the Chebychev inequality for which we assume that both parameters $\{p_a\}_{a\in A}$ are unknown. Then we make a study of the case when we know the value of p_1, but not that $p_1 > p_2$. Finally, working from the principle that if the agent does not implement action 2 very often, then for large n we can essentially reduce to the case when p_1 is known, we prove a tighter bound for the case when both arms are unknown through a contradiction argument.

Chebychev Bound Note that

$$1 - X_n = \mathbb{P}\left(\hat{p}_{2,n} - \hat{p}_{1,n} > 0 \mid \mathcal{X}_n\right).$$

and using Chebychev's inequality we find that

$$\begin{aligned}
&\mathbb{P}\left(\hat{p}_{2,n} - \hat{p}_{1,n} > 0 \mid \mathcal{X}_n\right) \\
&\quad = \mathbb{P}\left(\hat{p}_{2,n} - \hat{p}_{1,n} - (\bar{p}_{2,n} - \bar{p}_{1,n}) > (\bar{p}_{1,n} - \bar{p}_{2,n}) \mid \mathcal{X}_n\right) \\
&\quad \leq \frac{1}{1 + k_n^2}
\end{aligned}$$

where

$$\begin{aligned}
k_n^2 &= \left(\bar{p}_{1,n} - \bar{p}_{2,n}\right)^2 Var\left(\hat{p}_{2,n} - \hat{p}_{1,n} \mid \mathcal{X}_n\right)^{-1} \\
&= \left(\bar{p}_{1,n} - \bar{p}_{2,n}\right)^2 \left(\frac{(s_{2,n}+1)(f_{2,n}+1)}{(t_{2,n}+1)^2(t_{2,n}+2)} + \frac{(s_{1,n}+1)(f_{1,n}+1)}{(t_{1,n}+1)^2(t_{1,n}+2)}\right)^{-1}.
\end{aligned}$$

So since, by the Strong Law of Large Numbers

$$\frac{(s_{a,n}+1)(f_{a,n}+1)}{(t_{a,n}+1)^2} \to p_a \text{ for } a = 1, 2,$$

we have from A1 that

$$\lim_{n\to\infty} \frac{k_n^2}{t_{2,n}} = \frac{(p_1 - p_2)^2}{p_2(1-p_2)} \text{ almost surely.}$$

Hence for all $\epsilon > 0$, for large enough n

$$1 - X_n \leq (1+\epsilon)\frac{p_2(1-p_2)}{(p_1-p_2)^2}\frac{1}{t_{n,2}} \text{ a.s..}$$

Now for each $\epsilon, \delta > 0$, from Lemma 4.5 we have for large enough n

$$\Delta\xi(n) \leq (1+\epsilon)\frac{p_2(1-p_2)}{(p_1-p_2)^2}\frac{1}{F_\delta(\xi(n))} \text{ a.s.}$$

and so we deduce the following result from Proposition 4.4:

Proposition 4.7 *Suppose that $\bar{\xi}_{\epsilon,\delta}$ is a function solving the initial value problem: $\bar{\xi}_{\epsilon,\delta}(0) = \xi(0) \in (0,1)$, and*

$$\bar{\xi}'_{\epsilon,\delta}(n) = (1+\epsilon)\frac{p_2(1-p_2)}{(p_1-p_2)^2}\frac{1}{F_\delta\left(\bar{\xi}_{\epsilon,\delta}(n)\right)}.$$

Then for all $\epsilon, \delta > 0$,

$$\xi(n) = O\left(\bar{\xi}_{\epsilon,\delta}(n)\right) \text{ a.s..}$$

Remark 4.8 *Note that Chebychev's inequality is not sufficient here to demonstrate that $\xi(n) = O(\ln n)$.*

1 Arm Known Suppose that p_1 is known. Then $g_{1,n}(x) = \delta_{p_1}(x)$ for all n, and so

$$X_n = \int_{q=0}^{p_1} dG_{2,n}(q) = B_{x_1}(s_{2,n}+1, f_{2,n}+1)$$

where $B_x(\cdot,\cdot)$ is the incomplete beta function.

Now X_n only changes its value when the agent implements action 2. Using Sanov's Inequality we can bound the upper tail of the Beta distributions and obtain:

$$(1-X_n) \leq e^{-D(\bar{p}_2,p_1)t_{2,n}}$$

So once again using Lemma 4.5 we find for $\delta > 0$ and $D(p_2,p_1) > \epsilon > 0$ that, for large enough n,

$$\Delta\xi(n) \leq e^{-(D(p_2,p_1)-\epsilon)F_\delta(\xi(n))} \tag{4.6}$$

Hence we deduce the following:

Proposition 4.9 $\xi(n) = O(\ln n)$.

Proof. The proof is identical to that of Theorem 4.6 except that we use (4.6) in place of (4.5). ■

2 Arms Unknown Assume, once again, that both p_1 and p_2 are unknown to the agent.

The intuition and outline for our argument here is as follows. We begin from the assumption that $\xi(n)$ is increasing quite quickly as n gets large. As a result of this assumption the agent cannot be employing action 2 very frequently. Hence in comparison to the posterior $G_{2,n}$, the posterior $G_{1,n}$ begins to act like the Dirac measure δ_{p_1}. This motivates us to try to approximate X_n for large n by substituting δ_{p_1} for $G_{1,n}$ (i.e. reducing to the case of "1 Arm Known" studied above). Eventually this will give us a difference equation in $\xi(n)$, whose solution will be shown, by applying Proposition 4.4, to violate our assumption that $\xi(n)$ is growing rapidly in n, and the result will follow.

We shall write

$$Y_n := \int_{q=0}^{\bar{p}_{1,n}} dG_{2,n}(q) = \mathbb{P}(\hat{p}_{2,n} \leq \bar{p}_{1,n} \mid \mathcal{X}_n)$$

and in the following Lemma we prove that when $\xi(n)$ is growing sufficiently quickly we can reduce to analysing the more simple process Y_n:

Lemma 4.10 *Suppose that for some $\delta > 0$, $\gamma > \frac{1}{2}$ and $p_1 - p_2 > \epsilon > 0$, for large enough n*

$$F_\delta(\xi(n)) \geq \frac{1+\gamma}{D(\bar{p}_{2,n} - \epsilon, \bar{p}_{2,n})} \ln n \ \ a.s.. \tag{4.7}$$

Then

$$(1 - X_n) - (1 - Y_n) = o(1 - X_n).$$

Proof. Let $p_1 - p_2 > \epsilon > 0$.

We begin by conditioning on the value of $\hat{p}_{1,n}$:

$$\begin{aligned}
&(1 - X_n) - (1 - Y_n) \\
&\quad = \mathbb{P}(\hat{p}_{2,n} > \hat{p}_{1,n} \mid \mathcal{X}_n) - \mathbb{P}(\hat{p}_{2,n} > \bar{p}_{1,n} \mid \mathcal{X}_n) \\
&\quad = \mathbb{P}(\hat{p}_{2,n} > \hat{p}_{1,n} \mid \{\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon\} \cap \mathcal{X}_n)\, \mathbb{P}(\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n) \\
&\qquad + \mathbb{P}(\hat{p}_{2,n} > \hat{p}_{1,n} \mid \{\hat{p}_{1,n} > \bar{p}_{1,n} - \epsilon\} \cap \mathcal{X}_n)\, \mathbb{P}(\hat{p}_{1,n} > \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n) \\
&\qquad\qquad - \mathbb{P}(\hat{p}_{2,n} > \bar{p}_{1,n} \mid \mathcal{X}_n) \\
&\quad \leq \mathbb{P}(\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n) \\
&\qquad + [\mathbb{P}(\hat{p}_{2,n} > \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n) - \mathbb{P}(\hat{p}_{2,n} > \bar{p}_{1,n} \mid \mathcal{X}_n)]
\end{aligned}$$

So we proceed by showing how to control the tail probability and the square-bracketed difference.

Using Sanov's Inequality we can bound the tail probability:

$$\begin{aligned}\mathbb{P}(\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n) &= \mathbb{P}((1-\hat{p}_{1,n}) \geq 1 - (\bar{p}_{1,n} - \epsilon) \mid \mathcal{X}_n) \\ &\leq e^{-D(1-\bar{p}_{1,n}, 1-(\bar{p}_{1,n}-\epsilon))t_{1,n}}\end{aligned}$$

Now $D(\cdot,\cdot)$ is continuous and $D(x,y) \geq 0$ for all $x, y \in [0,1]$ with equality iff $x = y$. So we deduce that $\left(\mathbb{P}\left(\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n\right)\right)_{n\geq 1}$ is a summable sequence, and hence, since $\xi(n) \to \infty$, we must have

$$\mathbb{P}\left(\hat{p}_{1,n} \leq \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n\right) = o\left(1 - X_n\right).$$

Now we consider the square bracketed difference. Note that for large enough n

$$\mathbb{P}\left(\hat{p}_{2,n} > \bar{p}_{1,n} - \epsilon \mid \mathcal{X}_n\right) - \mathbb{P}\left(\hat{p}_{2,n} > \bar{p}_{1,n} \mid \mathcal{X}_n\right) \leq \epsilon g_{2,n}\left(\bar{p}_{1,n} - \epsilon\right).$$

Now

$$g_{2,n}(x) = \frac{1}{B\left(s_{2,n}+1, f_{2,n}+1\right)} x^{s_{2,n}} (1-x)^{f_{2,n}}.$$

Using Stirling's approximation for the Gamma Function we find that

$$\begin{aligned}B\left(s_{2,n}+1, f_{2,n}+1\right) &\sim \frac{\left(s_{2,n}+1\right)^{s_{2,n}+\frac{1}{2}} \left(f_{2,n}+1\right)^{f_{2,n}+\frac{1}{2}}}{\left(t_{2,n}+2\right)^{t_{2,n}+\frac{3}{2}}} \\ &= \left[\bar{p}_{2,n}\left(1-\bar{p}_{2,n}\right)\left(t_{2,n}+2\right)^{-1}\right]^{\frac{1}{2}} \left[\bar{p}_{2,n}^{\bar{p}_{2,n}} \left(1-\bar{p}_{2,n}\right)^{(1-\bar{p}_{2,n})}\right]^{t_{2,n}+2}.\end{aligned}$$

So

$$g_{2,n}(x) \sim \frac{\left(t_{2,n}+2\right)^{\frac{1}{2}}}{\sqrt{\bar{p}_{2,n}\left(1-\bar{p}_{2,n}\right)} x\left(1-x\right)} \left[\left(\frac{x}{\bar{p}_{2,n}}\right)^{\bar{p}_{2,n}} \left(\frac{1-x}{1-\bar{p}_{2,n}}\right)^{1-\bar{p}_{2,n}}\right]^{(t_{2,n}+2)}.$$

Now, for $x \in (0,1)$,

$$\begin{aligned}\frac{d}{dx}\left[x^p (1-x)^{1-p}\right] = p\left(\frac{1-x}{x}\right)^{1-p} - (1-p)\left(\frac{x}{1-x}\right)^{p} &= 0 \\ \iff x &= p.\end{aligned}$$

Hence, when $1 > x > \bar{p}_{2,n} > 0$,

$$0 < \left(\frac{x}{\bar{p}_{2,n}}\right)^{\bar{p}_{2,n}} \left(\frac{1-x}{1-\bar{p}_{2,n}}\right)^{1-\bar{p}_{2,n}} < 1,$$

and so, for large enough n,

$$g_{2,n}\left(\bar{p}_{1,n} - \epsilon\right) \sim \frac{t_{2,n}^{\frac{1}{2}} e^{-D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})t_{2,n}}}{\sqrt{\bar{p}_{2,n}\left(1-\bar{p}_{2,n}\right)}\left(\bar{p}_{1,n}-\epsilon\right)\left(1-\left(\bar{p}_{1,n}-\epsilon\right)\right)}.$$

Now suppose that for some $\gamma > \frac{1}{2}$, for all large enough n

$$F_\delta\left(\xi\left(n\right)\right) \geq \frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})} \ln n.$$

Then using the LIL, for all large enough n,

$$\begin{aligned}
&\frac{t_{2,n}^{\frac{1}{2}} e^{-D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})t_{2,n}}}{\sqrt{\bar{p}_{2,n}\left(1-\bar{p}_{2,n}\right)}\left(\bar{p}_{1,n}-\epsilon\right)\left(1-\left(\bar{p}_{1,n}-\epsilon\right)\right)} \\
&\quad \leq \frac{n^{\frac{1}{2}}}{\sqrt{\bar{p}_{2,n}\left(1-\bar{p}_{2,n}\right)}\left(\bar{p}_{1,n}-\epsilon\right)\left(1-\left(\bar{p}_{1,n}-\epsilon\right)\right)} \frac{1}{n^{1+\gamma}}.
\end{aligned}$$

Since the RHS of this last display is summable this allows us to conclude. ■

If we assume (4.7) holds then Lemma 4.10 tells us that

$$\begin{aligned}
\left(1-X_n\right) &= \left(1-Y_n\right) - \left[\left(1-Y_n\right) - \left(1-X_n\right)\right] \\
&= \left(1-Y_n\right) - o\left(1-X_n\right).
\end{aligned}$$

Now as before we may use Sanov's Inequality to bound the tail probability $(1-Y_n)$ and find that for all $p_1 - p_2 > \epsilon > 0$, for large enough n

$$\Delta\xi\left(n\right) \leq e^{-(D(p_2,p_1-\epsilon)-\epsilon)F_\delta(\xi(n))} + o\left(\Delta\xi\left(n\right)\right)$$

Our aim is to be able to use Proposition 4.4 to deduce from this difference equation an upper bound on the growth of ξ which contradicts our assumption (4.7) that ξ is growing quickly. The following Lemma is the technical result we need to make this argument work:

Lemma 4.11 *Let $C, \epsilon > 0$, $\gamma \geq \frac{1}{2} > \delta > 0$, and $f_{\delta,C}$ be the solution of the initial value problem:*

$$f_{\delta,C}(0) = 1 - X_0 \in (0,1), \; f'_{\delta,C}(n) = e^{-CF_\delta\left(f_{\delta,C}(n)\right)}. \tag{4.8}$$

Then, for any constant $c \geq 1$, the function $cf_{\delta,C}$ cannot satisfy (4.7) when for all large enough n

$$C\frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})} \geq 1.$$

Proof. Since the equation (4.8) can be written in the form $f'(n) = F(f(n))$ where F is uniformly Lipschitz on $[0,\infty)$ existence and uniqueness of the solution is guaranteed on $[0,\infty)$. Moreover, by considering the second derivative, we can deduce that $f_{\delta,C}$ is concave.

Let $\gamma > \frac{1}{2} > \epsilon > 0$, and suppose that for all large enough n

$$F_\delta(cf_{\delta,C}(n)) \geq \frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\ln n$$

for some constant $c > 0$. Then since

$$\frac{F_\delta(f_{\delta,C}(n))}{F_\delta(cf_{\delta,C}(n))} = \frac{f_{\delta,C}(n)\left(1-(f_{\delta,C}(n))^{\delta-\frac{1}{2}}\right)}{cf_{\delta,C}(n)\left(1-(cf_{\delta,C}(n))^{\delta-\frac{1}{2}}\right)} \to \frac{1}{c}$$

as $n \to \infty$, for any $\epsilon > 0$, for large enough n,

$$F_\delta(f_{\delta,C}(n)) \geq \frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\ln n \tag{4.9}$$

Now

$$\left(F_\delta(f_{\delta,C}(n))\right)' = \left(1-\frac{\frac{1}{2}+\delta}{f_{\delta,C}(n)^{\frac{1}{2}-\delta}}\right)f'_{\delta,C}(n)$$

and so

$$\begin{aligned}\left(e^{CF_\delta\left(f_{\delta,C}(n)\right)}\right)' &= C\left(F_\delta(f_{\epsilon,\delta,A}(n))\right)' e^{CF_\delta\left(f_{\delta,C}(n)\right)} \\ &= C\left(1-\frac{\frac{1}{2}+\delta}{f_{\delta,C}(n)^{\frac{1}{2}-\delta}}\right)f'_{\delta,C}(n)\, e^{CF_\delta\left(f_{\epsilon,\delta,A}(n)\right)} \\ &= C\left(1-\frac{\frac{1}{2}+\delta}{f_{\delta,C}(n)^{\frac{1}{2}-\delta}}\right). \end{aligned} \tag{4.10}$$

where we have used that $f_{\delta,C}$ solves (4.8).

So from (4.9) and (4.10) we obtain that, for large enough n,

$$\begin{aligned} C\frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}n^{C\frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}-1} & \\ &= \left(e^{C\frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\ln n}\right)' \\ &\leq \left(e^{CF_{\delta}\left(f_{\epsilon,\delta,A}(n)\right)}\right)' \\ &= C\left(1-\frac{\frac{1}{2}+\delta}{f_{\delta,C}\left(n\right)^{\frac{1}{2}-\delta}}\right) \\ &< C \end{aligned} \tag{4.11}$$

So fixing γ so that, for all large enough n,

$$C\frac{1}{(c+\epsilon)}\frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\geq 1$$

we find, by a contradiction in inequality (4.11), that $cf_{\delta,C}$ cannot satisfy the assumption (4.7). ■

We can now conclude the following Theorem:

Theorem 4.12 *For all* $\delta > 0$

$$F_{\delta}\left(\xi\left(n\right)\right) = O\left(\ln n\right).$$

Proof. Fix $p_1 - p_2 > \epsilon > 0$ and assume that for each $\gamma > \frac{1}{2} > \delta > 0$, for all large enough n

$$F_{\delta}\left(\xi\left(n\right)\right) \geq \frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\ln n$$

Then by Lemma 4.10

$$(1-X_n)-(1-Y_n) = o\left(1-X_n\right)$$

and so

$$\xi\left(n\right) = \sum_{k=1}^{n}\left(1-Y_k\right)+o\left(\xi\left(n\right)\right).$$

From Sanov's Inequality and Lemma 4.5 we deduce that

$$\Delta\xi\left(n\right) \leq e^{-(D(p_2,p_1-\epsilon)-\epsilon)F_\delta(\xi(n))} + o\left(\Delta\xi\left(n\right)\right)$$

and so we obtain from Proposition 4.4 that

$$\xi\left(n\right) = O\left(f_{\delta,D(p_1,p_2-\epsilon)-\epsilon}(n)\right)$$

where $f_{\delta,D(p_1,p_2-\epsilon)-\epsilon}$ is as defined in Lemma 4.11. So in particular

$$\xi\left(n\right) \leq cf_{\delta,D(p_1,p_2-\epsilon)-\epsilon}(n)$$

for some finite (random) constant $c \geq 1$.

For each possible value of $c \geq 1$ there exists a value of $\gamma > \frac{1}{2}$ such that for large enough n

$$\frac{1+\gamma}{(c+\epsilon)}\frac{D(p_1,p_2-\epsilon)-\epsilon}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})} \geq 1$$

Thus Lemma 4.11 provides a contradiction with the assumption that, for large enough n, $F_\delta\left(\xi\left(n\right)\right) \geq \frac{1+\gamma}{D(\bar{p}_{2,n}-\epsilon,\bar{p}_{2,n})}\ln n$ for all values of $\gamma > \frac{1}{2}$.

We conclude that $F_\delta\left(\xi\left(n\right)\right) = O\left(\ln n\right)$. ■

4.4 Conclusion

In this chapter we have given a regret analysis for Thompson Strategy for 2-Armed Bandits through analysing the quantity $t_{2,n}$. In particular in the cases of Bernoulli or Uniform rewards we show that the cumulative regret is at worst logarithmic in n, however it remains to be shown whether it achieves the optimal constant demonstrated in [4]. Moreover a finite time analysis of the algorithm needs to be provided.

We note that in empirical studies not included here show that the Thompson Strategy performs better than state-of-the-art algorithms such as KL-UCB [3] which have optimal asymptotic behaviour in their cummulative regret and for which finite time analyses has been performed.

Our proof of consistency is simple and general, and it allows for an extension of the algorithm to Contextualised Bandits and also a modification of the Thompson Strategy, the Optimistic Thompson Strategy (see [5]). However a proof of logarithmically increasing cumulative regret under general conditions on the rewards remains an open problem.

4.5 Appendix

Proof of Proposition 4.4. For any $\xi(0) \in (0,1)$, F is uniformly Lipschitz on $[\xi(0), \infty)$, and so there exists a unique solution, f, to the initial value problem

$$f'(x) = F(f(x)), \; f(0) = \xi(0),$$

and this solution is concave and increasing.

Suppose that for all $n \geq k$

$$\xi(n+1) - \xi(n) \leq F(\xi(n)).$$

Then to prove that

$$\xi(n) = O(f(n))$$

it is sufficient to show that for all large enough c, for all $x \geq k$,

$$(cf)(x+1) - (cf)(x) > F(cf(x)). \tag{4.12}$$

Observe that (4.12) holds iff

$$\begin{aligned} \frac{F(cf(x))}{c} &< f(x+1) - f(x) \\ &\leq f'(x) \\ &= F(f(x)) \end{aligned}$$

where we have used the concavity of f. So (4.12) holds iff

$$c > \frac{F(cf(x))}{F(f(x))}.$$

Now F is decreasing and so (4.12) must hold whenever $c > 1$. Therefore, taking $c = \max\left\{2, \frac{\xi(k)}{f(k)}\right\}$ we deduce that $\xi(n) = O(f(n))$. ■

Bibliography

[1] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412:1832–1852, 2011.

[2] Evan Fisher. On the law of the iterated logarithm for martingales. *Annals of Probability*, 20:675 – 680, 1992.

[3] Aurlien Garvier and Olivier Capp. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT, Hungary*, 2011.

[4] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4 – 22, 1985.

[5] B. May, N. Korda, A. Lee, and D. Leslie. Optimistic bayesian sampling in contextual-bandit problems. Submitted to Annals of Applied Probability.

[6] W.R. Thompson. On the likelihood that one unkown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[7] D. Williams. *Probability with Martingales*. CUP, 1991.