OPTIMAL DATA UTILIZATION FOR GOAL-ORIENTED LEARNING

BY CHARLES WESLEY COWAN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Mathematics

Written under the direction of

Michael Katehakis

and approved by

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

Optimal Data Utilization for Goal-Oriented Learning

by Charles Wesley Cowan

Dissertation Director: Michael Katehakis

We are interested in the problem of utilizing collected data to inform and direct learning towards a stated goal. In this work, a controller is presented with a finite set of actions that may be sequentially (and repeatedly) taken towards the achievement of some goal. While the outcome of any action is stochastic, the result provides information about future results of that action, and potentially others. By following a rule or control policy, the controller wishes to sequentially take actions, collect information, and utilize it towards future action decisions, in such a way as to approach the stated goal.

In the first model, at least one action is ‘best’, and the goal is to identify and take such an action as frequently as possible. This requires learning the actions’ underlying dynamics based on repeated observations of the stochastic results of those actions; this encapsulates the classic ‘exploration vs exploitation’ dynamic, to test many actions, or to take only the action currently believed to be best. We derive asymptotic lower bounds on how effective any universally good policy can be, as a function of initial knowledge. Additionally, we define a generic control policy and conditions under which it is provably asymptotically optimal, and give a number of examples to illustrate the scope and application of the model.

In the second model, the goal is to maximize some utility of all actions taken, e.g., total expected rewards collected. Additionally, each action has an associated breaking or halting time, which if reached ends the control process. This again captures the ‘exploration vs exploitation’ dynamic, as the controller must balance the reward of any one action against the risk of halting and loss of opportunity for future rewards. As the goal depends on the actual results achieved, there is generally no single ‘best’ action as in the previous model. In many contexts, we derive a dynamic ‘action valuation’ scheme that gives rise to an optimal control policy.

Acknowledgements

The issues addressed in this work, how to think about learning, and how to turn knowledge into action, have interested me for a very long time, so I am very grateful for the opportunity I have had here at Rutgers over the last 6 years to research and explore them. Too many have contributed in too many ways to thank or acknowledge, but I must try:

My thanks and gratitude to the members of my committee, for the consideration and review of this thesis. I am very grateful to my advisor, Professor Katehakis, for giving me both the language to address these problems, and the opportunity and support to do math that mattered to me.

I must thank the Rutgers Department of Mathematics, the Rutgers Business School, and the National Science Foundation for providing the support that allowed me to do this work. I would additionally like to thank the Mathematics Department, in particular Professors Han and Ocone, for their patience and support while I looked for a topic that suited me.

I would like to thank all of my friends in the department as well, anyone who ever put up with me coming into their office out of the blue and talking through a problem with them even when it was not relevant to their interest or expertise.

I'd also like to thank Maureen Clausen, Katie Guarino, Patty Sykes, and Luz Kosar, for their patience and immeasurable practical support in navigating this process.

Additionally, I would like to thank Professor Tumulka and Professor Mischaikow for inviting to join and supporting me on their work for a time, which while clearly separate from the work herein, I always viewed philosophically as of a piece.

Lastly, I'd like to thank Professor Tom Bohman of CMU, who recommended Rutgers to me.

Dedication

To Irene.

To my family.

To the friends I made along the way.

To Herbert Robbins, who I hope would have found something of interest herein.

Table of Contents

Chapter 1

Introduction: Bandits, Models for Learning

In this work, we are interested in the problem of learning, specifically in terms of collecting and utilizing data towards the achievement of some stated goal in the presence of uncertainty. The primary model under consideration is a Multi-Armed Bandit model, in which a controller is presented with a finite set of actions, e.g. 'arms' or 'bandits', that may be sequentially (and repeatedly) taken, with the results of the actions taken to contribute towards some goal. In general, the outcome of taking any action, while random, provides information about the underlying state or future results of that action - and potentially other actions, depending on what is known about how the actions relate to each other. The problem the controller faces is in utilizing this collected data as effectively as possible towards her goal. We are therefore interested in i) the development of decision rules or control policies for determining which action to take given the data available to optimally contribute to the controller's goal, and ii) analyzing the performance and behavior of such policies.

We consider two primary models in this work. In the first model, the subject of Chapters 2 and 3, at least one action is thought of as 'best' in some regard, and the controller's goal is to identify and take such an action as frequently as possible. This requires that the controller learn each action's underlying dynamics through repeated use and observation, but only to the extent necessary to identify and utilize the best of the available actions; this is the classic 'exploration vs exploitation' dilemma, balancing testing multiple actions for the purpose of learning and discovery against focusing exclusively on the single action currently believed to be best. Chapter 2 presents results bounding how efficiently any policy can discover and utilize the 'best' actions as a function of i) what is initially known about the nature of the actions and ii) the definition of 'best', as set by the controller. Chapter 3 develops a family of generic decision rules that can be applied in a variety of contexts, and additionally gives a set of sufficient conditions under which such policies are optimal, i.e., achieving the efficiency bounds as laid out in Chapter 2. Additionally, Chapter 3 gives a number of examples illustrating a breadth of application of this model. The main results of Chapter 3, the control policy defined therein and the asymptotic optimality result of Theorem 5, represent generalizations of the work of Cowan and Katehakis in [13] to non-i.i.d. observation processes.

The second model, which we refer to as the Halting Bandits model, is the subject of Chapter 4. In this model, the controller's goal is to take actions in such a way as to maximize some specified utility of the results of all actions taken, e.g., total expected rewards collected. Additionally, we take each action as having an associated 'halting' or breaking point, past which no further action can be taken; we view this as representing some conclusion, either a failure or success, and wish to maximize total utility up to this point. This halting bandit model differs significantly from the previous model in an important way: In the first model, there was (at least) one action that was 'best' to take for all time, and the results of actions taken mattered only in terms of the information they provided about which action was best. However, in this halting bandit model, because the stated goal depends on the actual results achieved, there is no one universally 'best' action - the best action to take at any time may depend on what is currently known about each action and the risk of halting, and may in fact change over time. We develop a scheme by which any action may be 'dynamically valued' based on what is currently known, and show in a variety of contexts that an optimal decision rule may be given by always taking the action with the largest current dynamic value. The results of this chapter are related to and can be seen as a generalization and extension of the model considered by Cowan and Katehakis in [15], as well as providing both a dramatic simplification of and greater intuition for the underlying optimality results of that paper.

Chapter 2

A Bound On Efficient Learning

In this chapter, we consider the following problem: a controller is faced with a finite set of actions, which may be taken (repeatedly) over time - the classic example is levers or arms to be pulled. At least one of the arms is considered ‘best’ to pull, relative to some stated goal, and the controller wishes to pull this arm as frequently as possible over some time horizon. However, the controller initially cannot be certain which of the arms is best. The controller must learn from the results of successive arm-pulls which arm is best. This is the classic ‘exploration vs exploitation’ dilemma: to what extent ought the controller experiment with various arms and learn more about them, versus sticking with or exploiting the arm she currently believes to be best?

We consider the problem of maximizing the expected number of ‘best’ actions, i.e. ‘optimal’ arm-pulls, over a given time horizon, or equivalently minimizing the expected number of ‘mistakes’ or sub-optimal pulls taken over that period. In particular, we wish to consider policies for sequentially deciding which arm to pull based on data collected, that minimize the expected number of mistakes ‘universally’ - for any set of arms with which the controller might be confronted. The primary result of this chapter is that for any such policy, subject to reasonable constraints on the set of potential arms, the expected number of mistakes must grow at least logarithmically with the total number of pulls. This represents a bound on how quickly or efficiently (in terms of the number of pulls taken) the identity of the best arm can be learned, depending on what is initially known about the arms, and how they relate to each other.

2.1 Formulation and Prior Work

The controller faces N ($2 \leq N < \infty$) arms, arm i represented by a sequence of random variables $\{X_t^i\}_{t \geqslant 1}$ on some Borel space $\mathscr{X}$. The t-th time arm i is pulled, the controller observes the value or state X_t^i. In this way, successive pulls reveal successively more information about the arm processes. We define an arm-pulling policy π as a stochastic process $\{\pi(n)\}_{n \geqslant 1}$ where $\pi(n) = i$ indicates that for the n-th pull, the controller pulls arm i. Given a policy π, we define $T_\pi^i(n) = \sum_{t=1}^n \mathbb{1}\{\pi(t) = i\}$, denoting the total number of pulls of arm i in the first n periods. We take $\pi(n+1)$ as dependent only on information available to the controller through the n-th pull, i.e., the decision of which arm to pull cannot depend on the results of arm pulls that have not happened yet.

The controller wishes to learn, via the results of pulls performed, about the underlying stochastic dynamics governing the arm processes, which are unknown to the controller. We consider 'learning' here to be the ability to increasingly differentiate between alternative hypotheses. To that end, we consider $\mathscr{F}$ as a known family of probability laws on $\mathscr{X}^\infty$, taking $\mathscr{F}$ to be the set of all plausible or potential hypotheses that might govern a given arm-process. For each i, we take the 'true' law governing $\{X_t^i\}_{t \geqslant 1}$ to be some element $F_i \in \mathscr{F}$, and denote $\underline{F} = (F_1, \ldots, F_N) \in \mathscr{F}^N$ as a full (ordered) set of arm laws. While $\mathscr{F}^N$ represents the full universe of potential arm laws, we restrict consideration in the following way: the controller may have additional information about the underlying structure of the arms and how they relate to each other, in particular that the laws $F_1, \ldots, F_N$ satisfy some property P. We may then define the full universe of plausible arm laws under consideration as

$$\mathscr{F}_P^N = \{(F_1, \ldots, F_N) \in \mathscr{F}^N : F_1, \ldots, F_N \text{ satisfy } P\}. \tag{2.1}$$

In the case that the arms are unrelated to each other, we simply take P as being trivially satisfied, so $\mathscr{F}_P^N = \mathscr{F}^N$. For a given set of laws $\underline{F}$, we denote $\mathbb{P}_{\underline{F}}(A)$ as the probability of event A when the true underlying arm laws are given by $\underline{F}$. As the controller observes more terms from arm i, she should hope to be increasingly able to distinguish F_i in $\mathscr{F}$.

We define the 'best' arm in the following way: we equip $\mathscr{F}$ with a 'score functional' $s : \mathscr{F} \mapsto \mathbb{R}$, providing a (problem specific) means of ranking the probability laws in $\mathscr{F}$. The classical score functional of interest is the long term average, $s(F) = F\text{-}\lim_n (1/n)\sum_{t=1}^{n} Y_t$, but we might also consider alternative scoring, such as other measures like a limiting median, or frequency of exceeding a given threshold. For a given set of arm laws $\underline{F} = (F_1, \ldots, F_N)$, let $s^*(\underline{F}) = \max_i s(F_i)$ be the maximal score of any arm. We may then define the best or 'optimal' arms as the set $O(\underline{F})$ where for $i \in O(\underline{F})$, $s(F_i) = s^*(\underline{F})$. The expected total number of 'mistakes' or sub-optimal pulls out of n pulls for a given policy is then

$$M_{\pi}^{\underline{F}}(n) = \mathbb{E}_{\underline{F}}\left[\sum_{i \notin O(\underline{F})} T_{\pi}^{i}(n)\right]. \tag{2.2}$$

For a given set of laws $\underline{F}$, it is reasonable to try to find π to minimize the above quantity (relative to the total number of pulls n) - consider, for instance, a policy that only pulls the optimal arms of $\underline{F}$! Implementing such a policy, however, depends on the controller initially knowing $\underline{F}$. The controller will not initially know what set of arms she faces. We may then ask, are there policies that minimize mistakes 'universally' over $\mathscr{F}_P^N$, so that the controller may be confident in the performance of the policy regardless of the specific arms she faces?

To begin, the number of mistakes is certainly bound by the total number of pulls, yielding the inequality $M_{\pi}^{\underline{F}}(n) \leq n$ for any $\underline{F} \in \mathscr{F}_P^N$. Hence, the expected number of mistakes can grow no worse than linearly with the number of pulls, universally. It is easy enough to establish policies such that $M_{\pi}^{\underline{F}}(n) = O(n)$ for all $\underline{F}$: consider for instance a policy that pulls every arm equally often - in this case, mistakes accumulate at a linear rate. Our primary interest is therefore policies for which the mistake rate is universally *sub*-linear, i.e., $M_{\pi}^{\underline{F}}(n) = o(n)$ for all $\underline{F} \in \mathscr{F}_P^N$. In particular, we define a policy π as $\mathscr{F}_P^N$**-Uniformly Fast** if for all $\alpha > 0$,

$$M_{\pi}^{\underline{F}}(n) = o(n^{\alpha}) \text{ for all } \underline{F} \in \mathscr{F}_P^N. \tag{2.3}$$

That is, a policy is uniformly fast if for any plausible choice of arms, the expected mistakes by time n accumulate slower than any power of n.

In the remainder of this chapter, we look at the implications of a policy being uniformly fast. In particular, in order to achieve slow mistake accumulation for *all plausible* choices of arms, such a policy must pull every arm sufficiently many times - and in doing so incur at least some mistakes - to be reasonable sure of identifying the best arms. It will be shown that enforcing such a slow rate of mistake accumulation in fact ensures that mistakes will accumulate at least at a logarithmic rate. We derive this minimum rate, and its dependence on $\mathscr{F}$, P, and s in the next section.

2.1.1 Prior Work

The problem as formulated here represents an important generalization of past work. Historically, the interest has centered on viewing the arm processes as a sequence of real-valued rewards to be collected. Hence for a given policy π, the *value* of a policy over a time horizon of n may be defined as

$$V_{\pi}^{\underline{F}}(n) = \sum_{i=1}^{N} \mathbb{E}_{\underline{F}} \left[\sum_{t=1}^{T_{\pi}^{i}(n)} X_{t}^{i} \right]. \tag{2.4}$$

Taking the processes to be i.i.d., arm i having mean μ_i, the above can be simplified to

$$V_{\pi}^{\underline{F}}(n) = \sum_{i=1}^{N} \mu_i \, \mathbb{E}_{\underline{F}} \left[T_{\pi}^{i}(n) \right]. \tag{2.5}$$

Comparing the above value to an idealized ‘optimal’ policy that always pulls the arm with maximal mean $\mu^* = \max_i \ \mu_i$ motivates the definition of regret or expected loss,

$$R_{\pi}^{\underline{F}}(n) = \mu^* n - V_{\pi}^{\underline{F}}(n) = \sum_{i=1}^{N} (\mu^* - \mu_i) \mathbb{E}_{\underline{F}} \left[T_{\pi}^{i}(n) \right]. \tag{2.6}$$

The focus has generally been on establishing policies that minimize regret, as a proxy for maximizing value. The following examples all consider cases where there is no known shared structure between the arms: In the $N = 2$ case, Robbins in [60] constructed policies that achieved regret as $o(n)$ universally, by ‘playing the current winner’ except for a sparse sequence of ‘forced’ exploratory pulls. Lai and Robbins in [44] introduced the idea of ‘uniformly fast’ policies, for which regret grew slower than any power of n, universally over arm laws. In the case of $\mathscr{F}$ as a one parameter family of densities, they proved that under mild regularity conditions, for any uniformly fast policy it must be that $R_{\pi}^{\underline{F}}(n) \geqslant \Omega(\ln n)$ for all $\underline{F}$, where the order constant depends on the specific $\underline{F}$. This logarithmic lower bound was generalized to the multi-parameter case by Katehakis and Burnetas in [9]. In general, policies that perform well cannot perform too well, as they must pull sub-optimal arms sufficiently many times (and in doing so, incur regret) so as to correctly identify the best arm.

The work presented here represents an extension and generalization of these previous works. In particular: we consider more general arm processes, relaxing the i.i.d. requirement; introduce potential shared structure relating the arms to each other, potentially allowing for faster learning due to greater knowledge; and introduce the idea of the 'general' score functional, allowing this learning framework to be applied in a wider variety of contexts. A key point in this generalization is the switch from regret minimization to mistake minimization. The two models are related, however: viewing the 'score' of an arm under this regret minimization model as the expected value, $s(F) = F\text{-}\lim_k 1/k \sum_{t=1}^{k} Y_t = \mathbb{E}_F[Y] = \mu(F)$, regret can be seen as a weighted sum of expected 'mistakes' for each sub-optimal arm, weight given by the expected loss for that arm.

Note that in the model of regret minimization, pulling arms that are 'close' to the optimal arm incurs less of a cost than distinctly less-optimal arms. Generalizing to the idea of a 'general' score functional, however, it is not immediately clear in the model presented here what is 'lost' via a sub-optimal pull: what is lost via a sub-optimal pull if arms are ranked according to median, or have infinite expected values? Hence we focus on 'total mistakes' rather than a notion of loss (though this model of mistakes is effectively imposing a 0-1 loss function). This becomes a natural generalization of the mentioned previous work, based on the following observation: the core of these previous results on bounding regret was first and foremost bounds on the individual mistake rates for sub-optimal arms, which were then combined via a weighted sum to produce a bound on regret. In this way, these past results are somewhat subsumed in and extended by the results to follow. Additionally, many of the results to follow may be applied to bound any loss function taking the form of a positive linear combination of the expected mistakes from each sub-optimal arm.

Additionally, it is interesting to note that the approach taken in this work, one might call it 'best arm utilization', seems to sit at a midpoint between the classical goals of 'regret minimization' as outlined above, and 'best arm identification', in which the focus is not on minimizing loss but rather maximizing the probability of correctly identifying the optimal arm after some period. In this model of 'best arm utilization', we abandon the goal of knowing which arm is truly best for the assurance that we will pull it as frequently as possible.

2.2 Preliminary Results for General Stochastic Processes

We begin with some notation and initial results relating to distinguishing, based on data, between two hypothesized probability laws in $\mathscr{F}$. To begin, we consider a fixed measure λ on $\mathscr{X}$, with λ^k the natural product measure on $\mathscr{X}^k$. For $F \in \mathscr{F}$, we characterize F by a family of finite dimensional density functions f, such that if $\{Y_t\}_{t \geqslant 1} \sim F$, for any $k > 0$, $(Y_1, \ldots, Y_k)$ is distributed according to density $f(y_1, \ldots, y_k)$ over $\mathscr{X}^k$, relative to λ^k. Note, in the case that F represents an i.i.d. sequence, we have that $f(y_1, \ldots, y_k) = \prod_{t=1}^k f(y_t)$. In general, the central restriction on f is that it be consistent with respect to the finite dimensional marginal distributions, for instance satisfying the following for any $k > 0$,

$$f(y_1, y_2, \ldots, y_k) = \int_{\mathscr{X}} f(y_1, y_2, \ldots, y_k, y_{k+1}) \lambda(dy_{k+1}). \tag{2.7}$$

An important result of this property is the following:

Lemma 1 *For any $k > 0$, if $f(y_1, \ldots, y_k) = 0$, then for any $k' > k$, $f(y_1, \ldots, y_k, y_{k+1}, \ldots, y_{k'}) = 0$ for all $(y_{k+1}, \ldots, y_{k'}) \in \mathscr{X}^{k'-k}$, $\lambda^{k'-k}$-almost everywhere.*

Proof. Without loss of generality, we may take $k' = k+1$. Note that f is strictly non-negative. If $f(y_1, \ldots, y_k, y_{k+1})$ were positive for a set of y_{k+1}-values of positive measure (w.r.t. λ), then $\int_{\mathscr{X}} f(y_1, y_2, \ldots, y_k, y_{k+1}) \lambda(dy_{k+1}) > 0$. By the above remark on consistent marginals, we have $f(y_1, \ldots, y_k) > 0$, a contradiction.

The implication of this is that if a sequence of values $(y_1, \ldots, y_k)$ is 'unlikely' relative to f, i.e., f gives a density of 0 at that point, then any sequence of values that begins with that sub-sequence is also unlikely.

Let F and G be two probability laws in $\mathscr{F}$. We define the following quantity as useful for 'hypothesis testing', determining whether a given process might be governed by F or by G:

$$\mathbf{I}(F, G) = \lim_{k \to \infty} \frac{1}{k} \ln \left(\frac{f(Y_1, \ldots, Y_k)}{g(Y_1, \ldots, Y_k)} \right) \text{ where } \{Y_t\}_{t \geqslant 1} \sim F, \tag{2.8}$$

interpreting the above limit as almost sure. The quantity $\mathbf{I}$ represents the 'limiting average log likelihood ratio'. If such a limit does not exist almost surely, it is convenient to take $\mathbf{I}(F, G)$ as infinite, to maintain consistency of the results to follow in that case.

An important property of this limit is the following:

Lemma 2 *For probability laws F, G, it holds that $\mathbf{I}(F,G) \geqslant 0$.*

Proof. This result follows naturally from the following result, that

$$\liminf_n \frac{1}{n} \ln\left(\frac{f(Y_1,\ldots,Y_n)}{g(Y_1,\ldots,Y_n)}\right) \geqslant 0 \ (\mathbb{P}_F\text{-a.s.}). \tag{2.9}$$

To see this, observe that for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}_F\left(\frac{1}{n}\ln\left(\frac{f(Y_1,\ldots,Y_n)}{g(Y_1,\ldots,Y_n)}\right) + \varepsilon < 0\right) \\
&= \mathbb{P}_F\left(f(Y_1,\ldots,Y_n) < g(Y_1,\ldots,Y_n)e^{-\varepsilon n}\right) \\
&= \mathbb{E}_F\left[\mathbb{1}\left\{f(Y_1,\ldots,Y_n) < g(Y_1,\ldots,Y_n)e^{-\varepsilon n}\right\}\right] \\
&\leq \mathbb{E}_G\left[\mathbb{1}\left\{f(Y_1,\ldots,Y_n) < g(Y_1,\ldots,Y_n)e^{-\varepsilon n}\right\}e^{-\varepsilon n}\right] \\
&\leq \mathbb{E}_G\left[1e^{-\varepsilon n}\right] \\
&= e^{-\varepsilon n}.
\end{aligned} \tag{2.10}$$

It follows that

$$\sum_{n=1}^{\infty} \mathbb{P}_F\left(\frac{1}{n}\ln\left(\frac{f(Y_1,\ldots,Y_n)}{g(Y_1,\ldots,Y_n)}\right) + \varepsilon < 0\right) < \infty. \tag{2.11}$$

By the Borel-Cantelli lemma, we have that for any $\varepsilon > 0$, $(1/n)\ln(f(Y_1,\ldots,Y_n)/g(Y_1,\ldots,Y_n)) < -\varepsilon$ only finitely often, $\mathbb{P}_F$-almost surely. This verifies Eq. (2.9). It follows then in the case that the limit exists almost surely, it must be that $\mathbf{I}(F,G) \geqslant 0$. In the case that the limit does not exist, we take $\mathbf{I}(F,G)$ as infinite, in which case the claim is trivially true.

We observe that, trivially, $\mathbf{I}(F,G) = 0$ if $F = G$. The function $\mathbf{I}(F,G)$ can be thought of as measuring a similarity - at least, in the limit for non-i.i.d. processes - between the distributions defined by F and G; the more similar F and G are, the smaller $\mathbf{I}(F,G)$ should be. If given data is likely to be generated under F, but unlikely to occur under G, the larger the likelihood ratio for that data, and the larger the resulting limit $\mathbf{I}(F,G)$ (typically). In particular, $\mathbf{I}(F,G)$ can be thought of in terms of the rate at which F is distinguished from G as more observations from F are made.

The following property will be useful for the results to follow:

Lemma 3 *For probability laws F and G, if $\mathbf{I}(F,G) < \infty$, then taking $\{Y_t\}_{t\geqslant 1} \sim F$, for all k,*

$$\frac{1}{k}\ln\left(\frac{f(Y_1,\ldots,Y_k)}{g(Y_1,\ldots,Y_k)}\right) < \infty \ \mathbb{P}_F\text{-almost surely.} \tag{2.12}$$

Proof. The assumption that $\mathbf{I}(F,G) < \infty$ implies that the limit of the average log likelihood exists and is finite almost surely (relative to $\mathbb{P}_F$). Suppose that the conclusion is false, that for some finite k_0, $(1/k_0)\ln(f(Y_1,\ldots,Y_k)/g(Y_1,\ldots,Y_k))$ is infinite with non-zero probability (relative to $\mathbb{P}_F$). In that case, we have that $g(Y_1,\ldots,Y_{k_0}) = 0$ with non-trivial probability (relative to $\mathbb{P}_F$). However, Lemma 1 therefore implies that for any $k > k_0$, given such $Y_1,\ldots,Y_{k_0}$, $g(Y_1,\ldots,Y_{k_0},y_{k_0+1},\ldots,y_k) = 0$ for almost every choice of $(y_{k_0+1},\ldots,y_k)$ (relative to λ^{k-k_0}). This implies that, given such $Y_1,\ldots,Y_{k_0}$, $(1/k)\ln(f(Y_1,\ldots,Y_k)/g(Y_1,\ldots,Y_k))$ is infinite $\mathbb{P}_F$-almost surely, and hence given the nontrivial probability of such $Y_1,\ldots,Y_{k_0}$, $\mathbf{I}(F,G)$ is infinite with non-trivial probability (relative to $\mathbb{P}_F$). This contradicts the existence of a finite $\mathbf{I}(F,G)$.

We establish this general framework so that the results to follow will be as broadly applicable as possible; the one central restriction in applying the results of the following sections is the assumption that for any F,G in the set of potential arm hypotheses $\mathscr{F}$, the limit $\mathbf{I}(F,G)$ exists almost surely, i.e., is a non-random constant value, depending only on F and G. This is not a small requirement. However, many models of interest do satisfy this requirement:

- **$\mathscr{F}$ as i.i.d. processes:** In this case, F and G represent i.i.d. sequences, and $\mathbf{I}$ reduces to

$$\mathbf{I}(F,G) = \lim_k \frac{1}{k}\sum_{t=1}^{k} \ln\left(\frac{f(Y_t)}{g(Y_t)}\right), \tag{2.13}$$

  which may be computed via the Strong Law as $\mathbf{I}(F,G) = \mathbb{E}_F[\ln(f(Y)/g(Y))]$, the usual Kullback-Leibler divergence, when the limit exists.

- **$\mathscr{F}$ as finite state Markov chains:** Let $\mathscr{X}$ be some finite state space, and consider the set of ergodic Markov chains on this space. In this case, any $F \in \mathscr{F}$ can be represented as an initial distribution p_0^F on $\mathscr{X}$, and a transition matrix P^F. In this case, $\mathbf{I}$ reduces to

$$\begin{aligned}\mathbf{I}(F,G) &= \lim_k \frac{1}{k}\left(\ln\left(\frac{p_0^F(Y_1)}{p_0^G(Y_1)}\right) + \sum_{t=2}^{k}\ln\left(\frac{P^F(Y_t|Y_{t-1})}{P^G(Y_t|Y_{t-1})}\right)\right)\\ &= \lim_k \frac{1}{k}\sum_{t=2}^{k}\ln\left(\frac{P^F(Y_t|Y_{t-1})}{P^G(Y_t|Y_{t-1})}\right).\end{aligned} \tag{2.14}$$

  This can be computed explicitly, based on the ergodic property of Markov chains. For any F, let $\underline{w}^F$ be the limiting distribution vector on $\mathscr{X}$. The above can be given explicitly as

$$\mathbf{I}(F,G) = \sum_{x \in \mathscr{X}} \underline{w}_x^F \sum_{y \in \mathscr{X}} P^F(y|x) \ln\left(\frac{P^F(y|x)}{P^G(y|x)}\right). \tag{2.15}$$

We note additionally that this extends naturally to k-th order finite state Markov processes as well.

- $\mathscr{F}$ **as hidden Markov models:** The quantity $\mathbf{I}$ can also be shown to exist in the case of taking $\mathscr{F}$ as a family of hidden Markov models, under certain regularity conditions on $\mathscr{F}$ such as stationarity of the underlying Markov chain [46]. There is no convenient formula to give for $\mathbf{I}$, but the existence of it proves that the results to follow, as bounds on a controller's ability to learn, are relevant to a number of applications.

2.3 A Lower Bound on Expected Mistakes

We can view learning in terms of differentiating between hypotheses as data is collected, for instance comparing two competing hypotheses that the arms are governed by $\underline{F}$ or $\underline{G}$, for $\underline{F}, \underline{G} \in \mathscr{F}_P^N$. In particular however, learning in this framework is modulated by the overarching goal: pulling the optimal arms (given the true underlying distributions). The following theorem can be thought of as bounding the minimum number of pulls needed to discern between the 'correct' hypothesis $\underline{F}$, and an alternative hypothesis $\underline{G}$. The case when $s^*(\underline{G}) > s^*(\underline{F})$ is of particular importance - if the correct hypothesis is $\underline{F}$, but the controller believes it to be $\underline{G}$, maximizing $\underline{G}$-optimal pulls would incur many mistakes!

The primary restriction we introduce here is taking the score functional to satisfy the following: *for any* $F, G \in \mathscr{F}$, *if* $\mathbf{I}(F,G) = 0$, *then* $s(F) = s(G)$. Essentially, this restricts $\mathscr{F}$ and s to say that if two process laws are indistinguishable *in the limit*, then they have the same score; this is reasonable: if a given arm is to be 'best' for all time, the property of being best will frequently be asymptotic in nature. This is frequently not a serious restriction, and is often satisfied for score functionals of interest such as those considered herein.

Theorem 1 *Let* π *be* $\mathscr{F}_P^N$*-Uniformly Fast. For any* $\underline{F} \in \mathscr{F}_P^N$*, let* $\underline{G} = (G_1, \ldots, G_N) \in \mathscr{F}_P^N$ *satisfy the following: i)* $s^*(\underline{G}) > s^*(\underline{F})$*, and ii)* $G_i = F_i$ *for all* $i \in O(\underline{F})$*. For any such* $\underline{F}, \underline{G}$*, let* $D(\underline{F}, \underline{G})$

be the set of all i such that $F_i \neq G_i$. Then the following holds:

$$\liminf_n \frac{\mathbb{E}_{\underline{F}}\left[\sum_{i\in D(\underline{F},\underline{G})} T^i_\pi(n)\right]}{\ln n} \geqslant \frac{1}{\mathbf{I}(\underline{F},\underline{G})}, \tag{2.16}$$

where $\mathbf{I}(\underline{F},\underline{G}) = \sum_i \mathbf{I}(F_i, G_i)$.

The proof is given in Section 2.5.

By choosing $\underline{G}$ to maximize the lower bound, the above may be utilized to bound the expected number of $\underline{F}$-mistakes, asymptotically, as $\underline{G}$ and $\underline{F}$ are taken to differ only on the sub-optimal arms of $\underline{F}$. By choosing specific $\underline{G}$ to agree or disagree with $\underline{F}$ on certain arms, the mistake rate relative to specific groups of arms can be bound as well. In the subsections to follow, we apply these notions to various specific instances of arms of interest. Note that the more restrictive the condition P (and hence the smaller $\mathscr{F}^N_P$), the fewer $\underline{G}$ there are that satisfy the conditions of Theorem 1 for a given $\underline{F}$, which reduces the maximal lower bound in Eq. (2.16). This has the following satisfying intuition: the more that is initially known about the arms (through P), the fewer mistakes that need to be made in learning and utilizing the optimal arms.

It is worth noting, however, that on occasion feasible $\underline{G}$ fail to exist: consider an $\underline{F} \in \mathscr{F}^N_P$ that realizes the a maximal value of s, for instance, in which case there is no $\underline{G} \in \mathscr{F}^N_P$ that has a better score than $\underline{F}$. Such $\underline{F}$ are somewhat privileged, and learning can on occasion be performed *faster* than logarithmically in these contexts [45]. This will not generally be the case in the models considered here, however.

2.3.1 Unrelated Arms

In this subsection, we consider applying Theorem 1 to the case of trivial P, which is to say that the individual arms share no known relationship with each other. This recovers and extends many classical results. We have the following general result:

Theorem 2 *Let π be $\mathscr{F}^N$-Uniformly Fast. For any $\underline{F} \in \mathscr{F}^N$, the following holds whenever the infima are taken over non-empty sets:*

$$\liminf_n \frac{M^{\underline{F}}_\pi(n)}{\ln n} \geqslant \sum_{i \notin O(\underline{F})} \frac{1}{\inf_{G\in\mathscr{F}}\{\mathbf{I}(F_i, G) : s(G) > s^*(\underline{F})\}}. \tag{2.17}$$

Proof. The proof proceeds by bounding the mistakes relative to any given sub-optimal arm $i \notin O(\underline{F})$. For such an i, consider a $\underline{G} = (G_1, \ldots, G_N) \in \mathscr{F}^N$ where $G_j = F_j$ for $j \neq i$, and $s(G_i) > s^*(\underline{F})$. Applying the result of Theorem 1, we have that

$$\liminf_n \frac{\mathbb{E}_{\underline{F}}\left[T_\pi^i(n)\right]}{\ln n} \geqslant \frac{1}{\mathbf{I}(F_i, G_i)}. \tag{2.18}$$

Maximizing the above lower bound relative to feasible G_i, and summing the resulting bound over all sub-optimal arms $i \notin O(\underline{F})$, completes the proof.

Note, the above proof can be extended to put a lower bound on any linear combination (with positive coefficients) of the individual sub-optimal mistakes as well. Taking the underlying arm processes to be i.i.d., and the score functional to be the expected value, $s(F) = \mathbb{E}_F[X]$, we recover Eq. (2.47) from the above, that for any $\underline{F}$, for any sub-optimal i, maximizing the bound over feasible $\underline{G}$,

$$\liminf_n \frac{\mathbb{E}_{\underline{F}}\left[T_\pi^i(n)\right]}{\ln n} \geqslant \frac{1}{\inf_{G \in \mathscr{F}}\{\mathbf{I}(F_i, G) : \mu(G) > \mu^*(\underline{F})\}}. \tag{2.19}$$

Weighting the mistakes of each sub-optimal arm by the expected loss, $\mu^*(\underline{F}) - \mu(F_i)$, we can combine the above bounds to recover a non-parametric version of the lower bounds of [44, 9].

$$\liminf_n \frac{R_\pi^{\underline{F}}(n)}{\ln n} \geqslant \sum_{i \notin O(\underline{F})} \frac{\mu^*(\underline{F}) - \mu(F_i)}{\inf_{G \in \mathscr{F}}\{\mathbf{I}(F_i, G) : \mu(G) > \mu^*(\underline{F})\}}. \tag{2.20}$$

2.3.2 Related Arms

In this subsection, we consider applying Theorem 1 to the case of non-trivial P - when arms are known to share a relationship. In this case, knowledge gained about one arm (through pulling it) can be informative not only about the law underlying that arm itself, but also about the other arms. As a given pull may therefore be more informative than in the unrelated case, we generally expect to be able to require fewer mistakes (and hence, a smaller lower bound) than in the unrelated arm case.

Note, the proof in the unrelated arm case bounded the total number of mistakes by first bounding the number of mistakes relative to each individual sub-optimal arm, through the construction of hypotheses $\underline{G}$ that differed from $\underline{F}$ *only* on a given arm. For non-trivial P - especially very restrictive P - given $\underline{F}$, it may not be possible to construct such a $\underline{G}$. As such, the following theorem provides several potential lower bounds:

Theorem 3 *Let π be $\mathscr{F}_P^N$-Uniformly Fast. For any $\underline{F} \in \mathscr{F}_P^N$, the following holds whenever the infimum is taken over a non-empty set:*

$$\liminf_n \frac{M_\pi^{\underline{F}}(n)}{\ln n} \geqslant \frac{1}{\inf_{\underline{G} \in \mathscr{F}_P^N} \{\mathbf{I}(\underline{F}, \underline{G}) : s^*(\underline{G}) > s^*(\underline{F}),\ \forall i \in O(\underline{F}) : G_i = F_i\}}. \tag{2.21}$$

Additionally, for any set S of sub-optimal arms in $\underline{F}$, i.e., if $i \in S$ then $i \notin O(\underline{F})$, the following holds whenever the infimum is taken over a non-empty set:

$$\liminf_n \frac{\mathbb{E}_{\underline{F}}\left[\sum_{i \in S} T_\pi^i(n)\right]}{\ln n} \geqslant \frac{1}{\inf_{\underline{G} \in \mathscr{F}_P^N} \{\mathbf{I}(\underline{F}, \underline{G}) : s^*(\underline{G}) > s^*(\underline{F}),\ \forall i \notin S : G_i = F_i\}}. \tag{2.22}$$

As an application of the above, the following holds whenever the infima are taken over non-empty sets:

$$\liminf_n \frac{M_\pi^{\underline{F}}(n)}{\ln n} \geqslant \sum_{i \notin O(\underline{F})} \frac{1}{\inf_{\underline{G} \in \mathscr{F}_P^N} \{\mathbf{I}(\underline{F}, \underline{G}) : s^*(\underline{G}) > s^*(\underline{F}),\ \forall j \neq i : G_j = F_j\}}. \tag{2.23}$$

Proof. All three bounds are natural results of Theorem 1. In particular, noting that for any feasible $\underline{G}$ (in the manner of Theorem 1) we have for $i \in D(\underline{F}, \underline{G})$ that $i \notin O(\underline{F})$, we have $\liminf_n M_\pi^{\underline{F}}(n)/\ln n \geqslant 1/\mathbf{I}(\underline{F}, \underline{G})$. Maximizing this lower bound over feasible $\underline{G}$ produces Eq. (2.21).

However, it is not immediately clear that Eq. (2.21) should be expected to be tight, especially in comparison to Theorem 2, which bounds the mistakes due to each sub-optimal arm individually. As it is not immediately clear mistakes can be bound on a per-arm basis (due to the restrictions of P), Eq. (2.22) considers a set of sub-optimal arms S, and follows from Theorem 1 in the same manner as Eq. (2.21), restricting in this case to $\underline{G}$ that differ from $\underline{F}$ only on S.

The bound in Eq. (2.23) is perhaps the most optimistic of the three, and the most likely to not exist (due to the restrictions of P). It follows in the same manner as Theorem 2.

2.3.3 Examples

We consider three examples to illustrate the above bounds. In each case, the results of a given arm are taken to be i.i.d., with normal distributions. This is to say, we restrict to the case of $\mathscr{F}$ as the set of normal, i.i.d. processes, with finite means and variances. Note, $\mathscr{F}$ can be parameterized in terms of the underlying mean (μ) and variance (σ^2). In this subsection, we consider the score functional on $\mathscr{F}$ as the expected value of the underlying distribution, $s(F) = \mathbb{E}_F[X] = \mu_F$. When the dependence on $\underline{F}$ is clear, we denote the maximal score $s^*(\underline{F})$ as μ^*, and the expected value and variance of F_i as μ_i and σ_i^2, respectively. It is straightforward to show, based on the previous commentary for $\mathbf{I}$ on i.i.d. processes, that

$$\mathbf{I}(F,G) = \frac{(\mu_F - \mu_G)^2}{2\sigma_G^2} + \frac{1}{2}\left(\frac{\sigma_F^2}{\sigma_G^2} - \ln\left(\frac{\sigma_F^2}{\sigma_G^2}\right) - 1\right). \tag{2.24}$$

Given this formula, it follows that if $\mathbf{I}(F,G) = 0$, it must be that $\mu_F = \mu_G$, and hence $s(F) = s(G)$. We apply this model, and the previous bounds, to the following cases:

- **Normal Arms with Unknown Means and Unknown Variances:** In this case, the set of feasible arm hypotheses is the full unrestricted space $\mathscr{F}^N$, i.e., P is taken to be trivial. Applying the results of Theorem 2, the bound may be computed explicitly as:

$$\liminf_n \frac{M_\pi^{\underline{F}}(n)}{\ln n} \geqslant \sum_{i:\mu_i \neq \mu^*} \frac{2}{\ln\left(1 + \frac{(\mu^* - \mu_i)^2}{\sigma_i^2}\right)}. \tag{2.25}$$

  Note, this is effectively the ‘mistake’ version of the regret bound addressed in [12].

- **Normal Arms with Unknown Means and Known Variances:** In this case, the set of feasible arm hypotheses is restricted, so that the variance of each arm is known in advance, i.e., P is taken to be the condition that for each i, $\mathrm{Var}_{F_i}(X) = \sigma_i^2$ for some known constant σ_i^2. Note that as P does not express a relationship between the individual arms, this model can actually be thought of as an extension of the case of unrelated arms, taking each arm i as having its own family of plausible hypotheses $\mathscr{F}_i$. Alternately, the bound of Theorem 3, Eq. (2.23) applies here as well:

$$\liminf_n \frac{M_\pi^{\underline{F}}(n)}{\ln n} \geqslant \sum_{i:\mu_i \neq \mu^*} \frac{2\sigma_i^2}{(\mu^* - \mu_i)^2}. \tag{2.26}$$

  Note, this is effectively the ‘mistake’ version of the regret bound addressed in [41].

- **Normal Arms with Unknown Means and Unknown, but Common, Variances:** In this case, the set of feasible arm hypotheses is restricted, so that the variance of each arm must be equal, i.e., P is taken to be the condition that for any feasible $\underline{F}$, $\mathrm{Var}_{F_1}(X) = \mathrm{Var}_{F_2}(X) = \ldots = \mathrm{Var}_{F_N}(X)$. Unlike the previous example, this condition explicitly relates each arm to the other. Because the means (and hence the potential scores) are unrestricted, however, the bound of Theorem 3, Eq. (2.23) may again be computed explicitly:

$$\liminf_{n} \frac{M_{\pi}^{\underline{F}}(n)}{\ln n} \geqslant \sum_{i:\mu_i \neq \mu^*} \frac{2\sigma^2}{(\mu^* - \mu_i)^2}, \tag{2.27}$$

where σ^2 is understood to be the common variance of $\underline{F}$.

Note that for a given $\underline{F}$, the bound in the case of unknown variances is *strictly* greater than or equal to the bound in the case of known variances. The greater knowledge afforded by the known variances potentially reduces the expected number of mistakes that a uniformly fast policy must incur. Further, it is interesting to note that the lower bound in the case of unknown, but common, variance is equal to that for the case of known variances, taking the known variances as equal. Simply knowing that the variance is common between arms is potentially (and in fact, will be shown to be in the next chapter) as useful in minimizing expected mistakes (asymptotically) as knowing what the variance actually is.

These examples will be explored in greater detail in the next chapter, where these bounds are shown to be tight.

2.4 When Efficient Learning is Impossible

The results of the previous sections, in particular Theorems 2, 3, demonstrate that for a universally good policy (i.e., uniformly fast), mistakes must accumulate at least logarithmically with the number of pulls, i.e., mistakes happen *at best* exponentially rarely over time. Such a policy is said to be *efficient*, efficiently learning and utilizing the best of the available arms.

However, there are models (choices of $\mathscr{F}, P, s$) such that efficient learning is impossible. This frequently arises in the following context: when, for a given set of arms, no matter how much data is accumulated about a sub-optimal arm through pulling, there always remain plausible alternative arm hypotheses which would make that arm the best arm of the set. The controller is essentially never willing to give up focus on an arm, no matter how poor the results of pulling it, for fear that at any moment it might yield something (however unlikely) that would make it the best arm. In such a case, mistakes accumulate strictly worse than logarithmically with the number of pulls. We have the following result,

Proposition 1 *There exist models (choices of $\mathscr{F}, P, s$), such that for any $\mathscr{F}_P^N$-Uniformly Fast policy π, for any $\underline{F} \in \mathscr{F}_P^N$, $\lim_n M_\pi^{\underline{F}}(n)/\ln n = \infty$.*

The proof is given in Section 2.5.

Analytically, the failure of efficient learning can be seen as caused in the following way: when for a given $F \in \mathscr{F}$, there are $G \in \mathscr{F}$ that are arbitrarily similar to F but arbitrarily better than F, i.e., $\mathbf{I}(F,G)$ is very small, but $s(G)$ is much larger than $s(F)$. This motivates the idea of imposing a continuity restriction on s relative to $\mathbf{I}$ over $\mathscr{F}$. While $\mathbf{I}$ is generally not a true metric, we can define a notion of continuity in the following way:

Definition 1 *A functional $s : \mathscr{F} \mapsto \mathbb{R}$ is continuous relative to $\mathbf{I}$ if for every $F \in \mathscr{F}$, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for any $G \in \mathscr{F}$, $\mathbf{I}(F,G) < \delta$ implies $|s(F) - s(G)| < \varepsilon$.*

Note, this definition of continuity is not symmetric with regards to the arguments of $\mathbf{I}$; in the definition of $\mathbf{I}$, the first argument is somewhat privileged, as it is taken to be the 'true' underlying probability law.

Restricting s to be continuous relative to $\mathbf{I}$ ensures that as more information is gained about the underlying arm laws, the controller may be increasingly sure that a given arm is sub-optimal. This notion of continuity is frequently satisfied by the score functionals of interest, such as the ones considered in this work.

2.5 Proofs

Proof. [of Theorem 1.] Let $\underline{F} \in \mathscr{F}$, and $\underline{G} \in \mathscr{F}$ be as hypothesized. Again, $D(\underline{F},\underline{G})$ is the set of i such that $F_i \neq G_i$. Note, the above is trivially true if $\mathbf{I}(\underline{F},\underline{G}) = \infty$, hence we may assume $\mathbf{I}(\underline{F},\underline{G})$ is finite, and therefore $\mathbf{I}(F_i,G_i) < \infty$ for each i. Additionally, the restriction on the score functional implies that since $s^*(\underline{G}) > s^*(\underline{F})$, we have that $\mathbf{I}(\underline{F},\underline{G}) > 0$, and $\mathbf{I}(F_i,G_i) > 0$ for some i.

Noting that

$$\mathbb{E}_{\underline{F}}\left[\left(\sum_{i \in D(\underline{F},\underline{G})} T_\pi^i(n)\right)\mathbf{I}(\underline{F},\underline{G})\right]/\ln n \geqslant \mathbb{P}_{\underline{F}}\left(\left(\sum_{i \in D(\underline{F},\underline{G})} T_\pi^i(n)\right)\mathbf{I}(\underline{F},\underline{G}) \geqslant \ln n\right), \quad (2.28)$$

it would suffice to show that

$$\liminf_n \mathbb{P}_{\underline{F}}\left(\frac{\sum_{i\in D(\underline{F},\underline{G})} T_\pi^i(n)}{\ln n} \geqslant \frac{1}{\mathbf{I}(\underline{F},\underline{G})}\right) = 1, \tag{2.29}$$

or equivalently that for $0<\delta<1$,

$$\limsup_n \mathbb{P}_{\underline{F}}\left(\frac{\sum_{i\in D(\underline{F},\underline{G})} T_\pi^i(n)}{\ln n} \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})}\right) = 0. \tag{2.30}$$

Define the following events:

$$A_n^\delta = \left\{ \sum_{i\in D(\underline{F},\underline{G})} T_\pi^i(n) \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})} \ln n \right\}, \tag{2.31}$$

$$C_n^\delta = \left\{ \sum_{i\in D(\underline{F},\underline{G})} \ln\left(\frac{f_i\left(X_1^i,\ldots,X_{T_\pi^i(n)}^i\right)}{g_i\left(X_1^i,\ldots,X_{T_\pi^i(n)}^i\right)} \right) \leq (1-\delta/2)\ln n \right\}. \tag{2.32}$$

It is additionally convenient to define the sequence of constants $b_n = (1-\delta)/\mathbf{I}(\underline{F},\underline{G})\ln n$ and random variables $S_k^i = \ln\left(f_i(X_1^i,\ldots,X_k^i)/g_i(X_1^i,\ldots,X_k^i)\right)$. Note that for $i\in D(\underline{F},\underline{G})$, we have in the case of A_n^δ, $T_\pi^i(n,) \leq b_n$. Hence we have the following bounds:

$$\begin{aligned}
\mathbb{P}_{\underline{F}}\left(A_n^\delta \bar{C}_n^\delta\right) &\leq \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})} \max_{k\leq\lfloor b_n\rfloor} S_k^i > (1-\delta/2)\ln n \right) \\
&= \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})} \max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > (1-\delta/2)\ln n/b_n \right) \\
&= \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})} \max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \left(1+\frac{\delta/2}{1-\delta}\right)\mathbf{I}(\underline{F},\underline{G}) \right) \\
&\leq \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})} \max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \left(1+\frac{\delta}{2}\right)\mathbf{I}(\underline{F},\underline{G}) \right).
\end{aligned} \tag{2.33}$$

Let $D_0 \subset D(\underline{F},\underline{G})$ such that for $i\in D_0$, $\mathbf{I}(F_i,G_i)=0$. Note, $\mathbf{I}(\underline{\mathrm{F}},\underline{\mathrm{G}}) = \sum_{i\in D(\underline{F},\underline{G})\setminus D_0}\mathbf{I}(F_i,G_i)$. The set D_0 may be empty, but we have that $|D_0|\leq N$. From the above,

$$\begin{aligned}
\mathbb{P}_{\underline{F}}\left(A_n^\delta \bar{C}_n^\delta\right) &\leq \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})\setminus D_0} \max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \left(1+\frac{\delta}{4}\right)\mathbf{I}(\underline{F},\underline{G}) \right) \\
&\quad + \mathbb{P}_{\underline{F}}\left(\sum_{i\in D_0} \max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \frac{\delta}{4}\mathbf{I}(\underline{F},\underline{G}) \right) \\
&\leq \sum_{i\in D(\underline{F},\underline{G})\setminus D_0} \mathbb{P}_{F_i}\left(\max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \left(1+\frac{\delta}{4}\right)\mathbf{I}(F_i,G_i) \right) \\
&\quad + \sum_{i\in D_0} \mathbb{P}_{F_i}\left(\max_{k\leq\lfloor b_n\rfloor} S_k^i/b_n > \frac{\delta}{4N}\mathbf{I}(\underline{F},\underline{G}) \right).
\end{aligned} \tag{2.34}$$

We may now make use of the following result: note that for $i \in D(\underline{F},\underline{G}) \setminus D_0$, $S^i_m/m \to \mathbf{I}(F_i,G_i)$ almost surely (relative to F_i), and for $i \in D_0$, $S^i_m/m \to 0$ almost surely (relative to F_i). We may then utilize the following proposition, the proof of which is given following the conclusion of this proof:

Proposition 2 *Let $\{S_m\}_{m\geqslant 1}$ be a sequence of almost surely finite random variables such that $S_m/m \to \mu$ almost surely, with $\mu \geqslant 0$. In that case, $\max_{k\leq m} S_k/m \to \mu$ almost surely as well.*

Note, we have that $\mathbf{I}(F_i,G_i) < \infty$, hence by Lemma 3, we have that the S^i_m are finite almost surely. Applying this to the above, we have that $\max_{k\leq m} S^i_k/m$ converges almost surely to $\mathbf{I}(F_i,G_i)$, which implies convergence in probability, and (recalling that $\mathbf{I}(\underline{F},\underline{G}) > 0$) yields the following bound:

$$\begin{aligned}\limsup_m \mathbb{P}_{\underline{F}}\left(A^\delta_m \bar{C}^\delta_m\right) &\leq \sum_{i\in D(\underline{F},\underline{G})\setminus D_0} \limsup_m \mathbb{P}_{F_i}\left(\max_{k\leq m} S^i_k/m > \left(1+\frac{\delta}{4}\right)\mathbf{I}(F_i,G_i)\right)\\ &+ \sum_{i\in D_0} \limsup_m \mathbb{P}_{F_i}\left(\max_{k\leq m} S^i_k/m > \frac{\delta}{4N}\mathbf{I}(\underline{F},\underline{G})\right)\\ &= 0.\end{aligned} \tag{2.35}$$

At this point, recall that $\mathbb{P}_{\underline{F}}$ has been defined by the choice of arm distributions $\underline{F} \in \mathscr{F}$. Consider defining the underlying probability space relative to $\underline{G}$ instead, with common probability measure $\mathbb{P}_{\underline{G}}$. The following holds:

$$\begin{aligned}&\mathbb{P}_{\underline{F}}\left(A^\delta_n C^\delta_n\right)\\ &= \mathbb{P}_{\underline{F}}\left(\sum_{i\in D(\underline{F},\underline{G})} T^i_\pi(n) \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})}\ln n,\ \prod_{i\in D(\underline{F},\underline{G})} \frac{f_i\left(X^i_1,\ldots,X^i_{T^i_\pi(n)}\right)}{g_i\left(X^i_1,\ldots,X^i_{T^i_\pi(n)}\right)} \leq n^{1-\delta/2}\right)\\ &\leq \mathbb{P}_{\underline{G}}\left(\sum_{i\in D(\underline{F},\underline{G})} T^i_\pi(n) \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})}\ln n\right) n^{1-\delta/2}.\end{aligned} \tag{2.36}$$

This change of measure argument follows, as C^δ_n restricts the region of probability space of interest to that where the comparison of laws that differ between $\underline{F}$ and $\underline{G}$ hold, i.e., $\prod_i f_i \leq n^{1-\delta/2}\prod_i g_i$. Note that since $s^*(\underline{G}) > s^*(\underline{F})$ and $\underline{F}$ and $\underline{G}$ agree on $O(\underline{F})$, we have that $O(\underline{G}) \subset D(\underline{F},\underline{G})$, hence

$$\mathbb{P}_{\underline{F}}\left(A^\delta_n C^\delta_n\right) \leq \mathbb{P}_{\underline{G}}\left(\sum_{i\in O(\underline{G})} T^i_\pi(n) \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})}\ln n\right) n^{1-\delta/2}, \tag{2.37}$$

or

$$\mathbb{P}_{\underline{F}}\left(A_n^{\delta} C_n^{\delta}\right) \leq \mathbb{P}_{\underline{G}}\left(n - \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})} \ln n \leq \sum_{i \notin O(\underline{G})} T_{\pi}^{i}(n)\right) n^{1-\delta/2}. \tag{2.38}$$

For n sufficiently large, so that $n > (1-\delta)/\mathbf{I}(\underline{F},\underline{G}) \ln n$, we may apply Markov's inequality to the above:

$$\mathbb{P}_{\underline{F}}\left(A_n^{\delta} C_n^{\delta}\right) \leq \frac{\mathbb{E}_{\underline{G}}\left[\sum_{i \notin O(\underline{G})} T_{\pi}^{i}(n)\right]}{n - \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})} \ln n} n^{1-\delta/2} = \frac{\mathbb{E}_{\underline{G}}\left[\sum_{i \notin O(\underline{G})} T_{\pi}^{i}(n)\right] n^{-\delta/2}}{1 - \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})} \frac{\ln n}{n}}. \tag{2.39}$$

Observing that under the assumption that π is $\mathscr{F}_N^P$-UF, $\mathbb{E}_{\underline{G}}[\sum_{i \notin O(\underline{G})} T_{\pi}^{i}(n)] = o(n^{\delta/2})$, it follows from the above that $\limsup_n \mathbb{P}_{\underline{F}}\left(A_n^{\delta} C_n^{\delta}\right) = 0$. Hence,

$$\begin{aligned} \limsup_{n} \mathbb{P}_{\underline{F}}\left(\frac{\sum_{i \in D(\underline{F},\underline{G})} T_{\pi}^{i}(n)}{\ln n} \leq \frac{1-\delta}{\mathbf{I}(\underline{F},\underline{G})}\right) \\ \leq \limsup_{n} \mathbb{P}\left(A_n^{\delta} C_n^{\delta}\right) + \limsup_{n} \mathbb{P}\left(A_n^{\delta} \bar{C}_n^{\delta}\right) = 0. \end{aligned} \tag{2.40}$$

Proof. [of Prop. 2] We have that S_m/m converges to $\mu \geqslant 0$ almost surely. First, we have that for $m \geqslant 1$,

$$\frac{S_m}{m} \leq \frac{\max_{k \leq m} S_k}{m}, \tag{2.41}$$

hence $\mu \leq \liminf_m \max_{k \leq m} S_k/m$.

Consider the case of $\mu > 0$: For a given m, define $k(m)$ to be the k that realizes $\max_{k \leq m} S_k$, so $\max_{k \leq m} S_k/m = S_{k(m)}/m$. Note the obvious result, $0 < k(m) \leq m$ almost surely. The $k(m)$ sequence must exist for $m = 1, 2, \ldots$, and in fact must almost surely increase without bound as $m \to \infty$, otherwise $\limsup_m S_{k(m)}/m \leq 0 < \mu$, contradicting the previous $\liminf$ result. Since S_m/m converges to $\mu > 0$ almost surely, almost surely there exists some finite M for which S_m is positive for all $m \geqslant M$. For m such that $k(m) > M$, it then follows that

$$\frac{S_{k(m)}}{m} = \frac{S_{k(m)}}{k(m)} \frac{k(m)}{m} \leq \frac{S_{k(m)}}{k(m)}. \tag{2.42}$$

From the above, and that $S_m/m \to \mu$ almost surely, it follows that $\limsup_m \max_{k \leq m} S_k/m \leq \mu$ almost surely.

In the case that $\mu = 0$: Let $k(m)$ be defined as in the previous case. Again the $k(m)$ sequence must exist for $m = 1, 2, \ldots$. If the $k(m)$ sequence does not increase without bound, trivially

$S_{k(m)}/m \to 0$. If the $k(m)$ sequence does increase without bound, but $S_{k(m)} \leq 0$ for all m, then $\limsup_m S_{k(m)}/m \leq 0$. If the $k(m)$ sequence does increase without bound, but the previous case does not hold, then $S_{k(m)} > 0$ for all sufficiently large m, and the argument from the previous $\mu > 0$ case applies. In all cases, $\limsup_m S_{k(m)}/m \leq 0$.

Combining the lim sup and lim inf results complete the proof.

Proof. [of Proposition 1.] The proof proceeds by example. Define $\mathscr{S}$ as the set of all finite unions of finite intervals on $\mathbb{R}$,

$$\mathscr{S} = \left\{ \bigcup_{j=1}^{k} [a_j, b_j] : 0 < k < \infty,\ -\infty < a_1 < b_1 < \ldots < a_k < b_k < \infty \right\}. \tag{2.43}$$

For $S \in \mathscr{S}$, it is convenient to define $|S| = \sum_j (b_j - a_j)$ as the measure of S.

We may then take $\mathscr{F}$ as the set of i.i.d. process laws F_S, with underlying distribution Unif(S), for any $S \in \mathscr{S}$. It is convenient to define $f_S(x) = \mathbb{1}\{x \in S\}/|S|$, the uniform density over S. We take P as trivially satisfied, which is to say we consider the arms as unrelated to each other in any known way. Additionally, we define the score functional $s : \mathscr{F} \mapsto \mathbb{R}$ as the expected value,

$$s(F_S) = \mu(F_S) = \int_{\mathbb{R}} x f_S(x) dx. \tag{2.44}$$

Under such a model, we have the following result,

$$\mathbf{I}(F_S, F_T) = \begin{cases} \ln|T| - \ln|S| & \text{if } S \subset T \\ \infty & \text{else.} \end{cases} \tag{2.45}$$

Note, if $\mathbf{I}(F, G) = 0$, it implies that F and G have the same support (up to a set of measure 0), and therefore $\mu(F) = \mu(G)$.

For a given choice of arm laws $\underline{F} \in \mathscr{F}^N$, let S_i be the support of F_i, and let $\mu^* = \max_i \mu(F_i)$. For any sub-optimal i and $\varepsilon > 0$, let $\tilde{S}_i = S_i \cup I_\varepsilon$, where I_ε is an interval of width epsilon. For I_ε sufficiently far to the right, i.e., so that S_i does not intersect I_ε, we have that $\mu(F_{\tilde{S}_i}) > \mu^*$, and additionally that

$$\mathbf{I}(F_i, F_{\tilde{S}_i}) = \ln(|S_i| + \varepsilon) - \ln|S_i|. \tag{2.46}$$

We therefore have, applying Eq. (2.47)

$$\liminf_n \frac{\mathbb{E}_{\underline{F}}\left[T_\pi^i(n)\right]}{\ln n} \geqslant \frac{1}{\ln(|S_i| + \varepsilon) - \ln|S_i|}. \tag{2.47}$$

Taking the limit, as $\varepsilon \to 0$, we have that for any sub-optimal i, $\liminf_n \mathbb{E}_{\underline{F}}\left[T_\pi^i(n)\right]/\ln n = \infty$, i.e., the number of mistakes for arm i grows super-logarithmically in expectation. The result follows from this, applying it to each sub-optimal arm.

Chapter 3

UCB Policies for Maximizing Optimal Utilization

In the previous chapter, we derived limits on how efficiently ‘universally good’ policies can learn the best arm of a set of arms. This chapter builds on that premise, demonstrating that in many contexts these limits are ‘tight’, in the sense of being realized by implementable arm-pulling policies. We demonstrate such a policy, showing how the data collected from each arm can be utilized to assign a value or index to each arm, such that always pulling the arm with the current highest index is an asymptotically optimal policy, in the sense of the previous chapter. The primary focus in this chapter is on arms that are unrelated to each other, but the case of arms restricted by some common structure is also discussed.

3.1 Formulation and Prior Work

We adopt the notation of the previous chapter, that the controller faces N ($2 \leq N < \infty$) arms, each represented by a sequence $\{X^i_t\}_{t \geqslant 1}$ of random variables on some space $\mathscr{X}$, with underlying probability laws $\underline{F} = (F_1, \ldots, F_N) \in \mathscr{F}^N$. We focus primarily on the case of unrelated arms, i.e., the full set of plausible arm hypotheses is taken to be $\mathscr{F}^N$.

Given the results of the previous section, we are generally interested in establishing policies π such that for all $\underline{F} \in \mathscr{F}^N$, $M^{\underline{F}}_{\pi}(n) = O(\ln n)$. In particular, defining the function

$$\mathbb{K}_F(\rho) = \inf_{\tilde{F} \in \mathscr{F}} \{\mathbf{I}(F, \tilde{F}) : s(\tilde{F}) > \rho\}, \tag{3.1}$$

we are interested in policies that achieve the limit of Theorem 2, i.e.,

$$\lim_n \frac{M^{\underline{F}}_{\pi}(n)}{\ln n} = \sum_{i \notin O(\underline{F})} \frac{1}{\mathbb{K}_{F_i}(s^*(\underline{F}))}. \tag{3.2}$$

Such policies are referred to as **Asymptotically Optimal**.

3.1.1 Prior Work

As in the previous chapter, prior work in this area has focused on policies that minimize regret in the i.i.d., unrelated arm case, i.e., expected mistakes weighted according to expected loss,

$$R_{\pi}^{\underline{F}}(n) = \sum_{i \notin O(\underline{F})} (\mu^*(\underline{F}) - \mu(F_i)) \mathbb{E}_{\underline{F}} \left[T_{\pi}^{i}(n) \right]. \tag{3.3}$$

A variety of policies have been introduced to try to address the problem of regret minimization. Broadly, there are roughly two classes of techniques and effects that are generally utilized: upper confidence bound (UCB) index policies that compute a ranking of arms based on current data by placing a confidence interval on the true underlying expected value for an arm, and pull the highest ranked arm; and Bayesian (Thompson Sampling) policies that pull arms according to the posterior probabilities of which arm is best. While there has been some notable success in analyzing the behavior of Thompson Sampling policies [43, 36, 3], the focus of this current work and this current chapter is on a generalization of UCB index policies.

The results of prior work on regret minimization (e.g., Section 2.1.1) generally indicate that universally good policies can do no better than incurring logarithmic regret for any $\underline{F} \in \mathscr{F}^N$. As such, much effort has been spent demonstrating simple policies that satisfy $R_{\pi}^{\underline{F}}(n) = O(\ln n)$. For example, in [5], Auer et al. present a policy which is provably logarithmic in regret under very weak conditions on $\mathscr{F}$. However, as the result of the previous chapter and the related prior work indicate, such policies have an asymptotic lower bound on how small regret can be. Policies that achieve minimal regret asymptotically have been established in a variety of contexts, such as normal arms with known variances [41], normal arms with unknown variances [12, 36], and arms with multinomial returns [9, 35].

Generic policies, ones that might be applied in a variety of contexts, have also been established and proven asymptotically optimal under mild regularity conditions in the case of $\mathscr{F}$ as a one-parameter family of distributions [44], or multi-parameter / potentially mixed family of distributions [9]. The policy developed in [9] was independently developed and applied in [11].

The results presented in this work can in fact be seen as an extension and generalization of the UCB policy established in [9]; we introduce a correction to the policy presented there to render it provably asymptotically optimal in more contexts, as well as generalizing that policy to non-i.i.d. arm processes and generic score functions. Additionally, we generalize and expand the 'mild regularity conditions' of [9] on which the asymptotic optimality of the policy relies, allowing for both broader applicability, and simplified proofs of optimality.

3.2 An Asymptotically Optimal UCB Policy

In this section, we construct a policy, in the spirit of classical UCB policies, for determining which arm to pull next give the data currently available. The policy constructed here will generally perform well, but under certain conditions can be shown to be asymptotically optimal.

For a given $F \in \mathscr{F}$, let $\hat{F}_t = \hat{F}_t[X_1, \ldots, X_t] \in \mathscr{F}$ be an estimator of F given t samples from F. We make few restrictions on the nature of these estimators, except to say that $\hat{F}_t$ should converge to F with t, in a manner that will be made more clear shortly. To define convergence in $\mathscr{F}$, we need some notion of a distance in $\mathscr{F}$. While **I** can frequently serve as a similarity measure on $\mathscr{F}$, it is often convenient to consider alternative similarity measures. Let ν be a (context-specific) measure of similarity on $\mathscr{F}$; for instance, if $\mathscr{F}$ is parameterized, ν might be the ℓ_2-norm on the parameter space. We restrict $\mathscr{F}$, s, $\hat{F}_t$, ν, by assuming the following conditions hold, for any $F \in \mathscr{F}$, and all $\varepsilon, \delta > 0$:

- **Condition 1:** $\mathbb{K}_F(\rho)$ is continuous with respect to ρ, and with respect to F under ν (in the manner of Def. 1).
- **Condition 2:** $\mathbb{P}_F\left(\nu(\hat{F}_t, F) > \delta\right) \leq o(1/t)$.
- **Condition 3:** For some sequence $d_t = o(t) \geqslant 0$ (independent of ε, δ, F),

$$\mathbb{P}_F\left(\delta < \mathbb{K}_{\hat{F}_t}(s(F) - \varepsilon)\right) \leq e^{-\Omega(t)} e^{-(t-d_t)\delta}, \tag{3.4}$$

  where the dependence on ε and F are suppressed into the $\Omega(t)$ term.

Condition 1 in some sense characterizes the structure of $\mathscr{F}$ as smooth relative to s; while **I** is not a metric on $\mathscr{F}$, to the extent that it can be thought of as a measure of similarity, $\mathbb{K}_F(\rho)$ can be thought of as a Hausdorff distance on $\mathscr{F}$, and hence **Condition 1** additionally restricts the 'shape' of $\mathscr{F}$ relative to s. **Condition 2** in some sense merely states that the estimators

$\hat{F}_t$ are ‘honest’, and converge to F sufficiently quickly with t. **Condition 3** often seems to be satisfied by $\hat{F}_t$ converging to F sufficiently quickly, as well as $\hat{F}_t$ being ‘useful’, in that $s(\hat{F}_t)$ converges sufficiently quickly to $s(F)$ as well. The bound in **Condition 3**, while oddly specific in its dependence on t, δ, can be relaxed somewhat, but such a bound frequently seems to exist in practice, for natural choices of $\hat{F}_t$.

With these restrictions in mind, we define the following policy:

Let $\tilde{d}(t) > 0$ be a non-decreasing, sub-linear function. Define, for any t such that $t > \tilde{d}(t)$, the following index function:

$$u_i(n,t) = \sup_{G\in\mathscr{F}} \left\{ s(G) : \mathbf{I}(\hat{F}^i_t, G) < \frac{\ln n}{t - \tilde{d}(t)} \right\}. \tag{3.5}$$

For a given $\tilde{d}$, let $n_0 \geqslant \min\{n : n > \tilde{d}(n)\}$. We then define the following generic policy,

Policy π^*: UCB-$(\mathscr{F}, s, \tilde{d})$:

- i) For $n = 1, 2, \ldots, n_0 \times N$, pull each arm n_0 times to construct initial estimators,
- ii) For $n \geqslant n_0 \times N$, pull arm $\pi^*(n+1) = \text{argmax}_i\{u_i(n, T^i_{\pi^*}(n))\}$, breaking ties uniformly at random.

The following theorem characterizes the sub-optimal pulls of policy π^*:

Theorem 4 *Let $\underline{F} \in \mathscr{F}^N$ be a choice of arm laws. Under the above policy π^*, for any sub-optimal $i \notin O(\underline{F})$, and optimal $i^* \in O(\underline{F})$, the following result holds for any $\varepsilon > 0$ such that $s^*(\underline{F}) - \varepsilon > s(F_i)$, and $\delta > 0$ such that $\inf_{G\in\mathscr{F}}\{\mathbb{K}_G(s^*(\underline{F}) - \varepsilon) : \nu(G, F_i) \leq \delta\} > 0$:*

$$\begin{aligned} \mathbb{E}_{\underline{F}}\left[T^i_{\pi^*}(n)\right] \leq & \frac{\ln n}{\inf_{G\in\mathscr{F}}\{\mathbb{K}_G(s^*(\underline{F}) - \varepsilon) : \nu(G, F_i) \leq \delta\}} + o(\ln n) \\ & + \sum_{t=n_0N}^{n} \mathbb{P}_{F_i}\left(\nu(\hat{F}^i_t, F_i) > \delta\right) \\ & + \sum_{t=n_0N}^{n} \sum_{k=n_0}^{t} \mathbb{P}_{F_{i^*}}\left(u_{i^*}(t,k) \leq s^*(\underline{F}) - \varepsilon\right). \end{aligned} \tag{3.6}$$

The proof is given in Section 3.5.

The above holds generally, but when **Conditions 1-3** are additionally met, the above theorem can be utilized to demonstrate asymptotic optimality of π^*.

Theorem 5 *Let $\mathscr{F}$, s, $\hat{F}_t$, and ν satisfy **Conditions 1-3**, and additionally that s is continuous over $\mathscr{F}$ with respect to* $\mathbf{I}$*. Let d be as in **Condition 3**. If $\tilde{d}(t) - d_t \geqslant \Delta > 0$ for some Δ, for all t, then π^* is asymptotically optimal, i.e., for all $\underline{F} \in \mathscr{F}^N$ where the infima are defined,*

$$\lim_n \frac{M_\pi^{\underline{F}}(n)}{\ln n} = \sum_{i \notin O(\underline{F})} \frac{1}{\mathbb{K}_{F_i}(s^*(\underline{F}))}. \tag{3.7}$$

Proof. Consider a given $\underline{F} \in \mathscr{F}^N$, and let $i \notin O(\underline{F})$ be a sub-optimal arm, and $i^* \in O(\underline{F})$ be an optimal arm. There trivially exist feasible ε as in Theorem 4. By the continuity of s with respect to $\mathbf{I}$, $\mathbb{K}_F(\rho) > 0$ for all $\rho > s(F)$. It follows from this, and the continuity of $\mathbb{K}_F(\rho)$ with respect to F under ν that all sufficiently small $\delta > 0$ are feasible. Let ε, δ be feasible as in Theorem 4.

Note, by **Condition 2**,

$$\sum_{t=1}^{n} \mathbb{P}_{F_i}(\nu(\hat{F}_t^i, F_i) > \delta) \leq \sum_{t=1}^{n} o(1/t) \leq o(\ln n). \tag{3.8}$$

Similarly, by **Condition 3**, for $k \geqslant n_0$, (noting that $s(F_{i^*}) = s^*(\underline{F}) = s^*$),

$$\begin{aligned}
\mathbb{P}_{F_{i^*}}\left(u_{i^*}(t,k) \leq s^* - \varepsilon\right) &= \mathbb{P}_{F_{i^*}}\left(\sup_{G \in \mathscr{F}}\left\{s(G) : \mathbf{I}(\hat{F}_k^{i^*}, G) < \frac{\ln t}{k - \tilde{d}(k)}\right\} \leq s(F_{i^*}) - \varepsilon\right) \\
&\leq \mathbb{P}_{F_{i^*}}\left(\inf_{G \in \mathscr{F}}\left\{\mathbf{I}(\hat{F}_k^{i^*}, G) : s(G) > s(F_{i^*}) - \varepsilon\right\} > \frac{\ln t}{k - \tilde{d}(k)}\right) \\
&\leq e^{-\Omega(k)} e^{-(k-d(k))\frac{\ln t}{k-\tilde{d}(k)}} \\
&= \frac{1}{t} t^{-\frac{\tilde{d}(k)-d_k}{k-\tilde{d}(k)}} e^{-\Omega(k)} \\
&\leq \frac{1}{t} t^{-\frac{\Delta}{k-\tilde{d}(k)}} e^{-\Omega(k)}
\end{aligned} \tag{3.9}$$

Hence,

$$\sum_{k=n_0}^{t} \mathbb{P}_{F_{i^*}}\left(u_{i^*}(t,k) \leq s^*(\underline{F}) - \varepsilon\right) \leq \frac{1}{t} \sum_{k=1}^{\infty} t^{-\frac{\Delta}{k-\tilde{d}(k)}} e^{-\Omega(k)} \leq \frac{1}{t} O(1/\ln t). \tag{3.10}$$

The last step is proven as Proposition 3 in Section 3.5.

From Theorem 4,

$$\begin{aligned}
\mathbb{E}_{\underline{F}}\left[T_{\pi^*}^i(n)\right] &\leq \frac{\ln n}{\inf_{G \in \mathscr{F}}\{\mathbb{K}_G(s^*(\underline{F}) - \varepsilon) : \nu(G, F_i) \leq \delta\}} + \sum_{t=1}^{n} \frac{1}{t} O(1/\ln t) + o(\ln n) \\
&= \frac{\ln n}{\inf_{G \in \mathscr{F}}\{\mathbb{K}_G(s^*(\underline{F}) - \varepsilon) : \nu(G, F_i) \leq \delta\}} + O(\ln \ln n) + o(\ln n).
\end{aligned} \tag{3.11}$$

Hence it follows,

$$\limsup_n \frac{\mathbb{E}_{\underline{F}}\left[T_{\pi^*}^i(n)\right]}{\ln n} \leq \frac{1}{\inf_{G \in \mathscr{F}}\{\mathbb{K}_G(s^*(\underline{F}) - \varepsilon) : \nu(G, F_i) \leq \delta\}}. \tag{3.12}$$

By the continuity of $\mathbb{K}$ as under **Condition 1**, minimizing the above bound first with respect to δ, then ε, yields

$$\limsup_{n} \frac{\mathbb{E}_{\underline{F}}\left[T^{i}_{\pi^*}(n)\right]}{\ln n} \leq \frac{1}{\mathbb{K}_{F_i}(s^*(\underline{F}))}. \tag{3.13}$$

By the continuity of s with respect to $\mathbf{I}$, the lower bound on the limit is given with the lim inf via Theorem 2, hence for each sub-optimal i,

$$\lim_{n} \frac{\mathbb{E}_{\underline{F}}\left[T^{i}_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\mathbb{K}_{F_i}(s^*(\underline{F}))}. \tag{3.14}$$

Summing over sub-optimal i completes the result.

For specific contexts, i.e., choices of $\mathscr{F}$, s, verifying π^* as asymptotically optimal is reduced to finding $\hat{F}_t$, ν to verify **Conditions 1-3**, as well as verifying the continuity of s with respect to $\mathbf{I}$. The continuity of s, as well as **Conditions 1 & 2**, generally seem easy to verify, particularly when $\mathscr{F}$ is parameterized and those parameters can be efficiently estimated. The difficulty frequently lies in verifying **Condition 3** - verifying similar conditions have been the brunt of the work in verifying asymptotically optimal policies for regret minimization in many contexts [9, 12, 14, 36, 41]. However, advance knowledge of the specific form of the bound as given in **Condition 3** is frequently helpful in finding and verifying such a bound in practice.

Noting that Theorem 5 is essentially just an asymptotic upper bound on the results of Theorem 4, we observe that for specific contexts, the bound of Theorem 4 can often be computed more precisely, yielding finite horizon bounds, as well as estimates of the asymptotic remainder term on $M^{\underline{F}}_{\pi^*}(n)$. We do not focus on this in this work, however, and the remainder of this chapter is primarily devoted to demonstrating the asymptotic optimality of π^* in a variety of contexts. Section 3.3 is devoted to i.i.d., unrelated arm processes in various frameworks of interest, while Section 3.4 demonstrates the asymptotic optimality of a related UCB Index policy on an example with related arm processes.

3.3 Examples of Interest

In this section, we continue the focus on unrelated arms, i.e., taking P as trivial, so $\mathscr{F}^N$ represents the full space of plausible arm hypotheses. Additionally, we restrict to the case of i.i.d. arms, in which case $\mathscr{F}$ may be identified with the densities of the underlying i.i.d. processes, which we represent with a lower case f. The i.i.d. assumption is not necessary, as the results of Theorems 4, 5 hold in the general case, but many examples of interest arise in the i.i.d. case.

3.3.1 Pareto Arms with Separable Scores

In this section, we consider a model that demonstrates the utility of this generalized score functional approach. We take $\mathscr{F} = \mathscr{F}_\ell$, for $\ell \geqslant 0$, as the family of Pareto distributions defined by:

$$\mathscr{F}_\ell = \left\{ f_{\alpha,\beta}(x) = \frac{\alpha\beta^\alpha}{x^{1+\alpha}} : \alpha > \ell, \beta > 0 \right\}. \tag{3.15}$$

Taking X as distributed according to $f_{\alpha,\beta} \in \mathscr{F}_\ell$, e.g., $X \sim \text{Pareto}(\alpha,\beta)$, X is distributed over $[\beta,\infty)$, with $\mathbb{E}[X] = \alpha\beta/(\alpha-1)$ if $\alpha > 1$, and $\mathbb{E}[X]$ as infinite or undefined if $\alpha \leq 1$. We are particularly interested in $\mathscr{F}_0$, the family of unrestricted Pareto distributions, and $\mathscr{F}_1$, the family of Pareto distributions with finite means.

Under the general goal of obtaining large rewards from the arms pulled, there are two effects of interests: rewards from a given arm will be biased towards larger values for increasing β and decreasing α. Hence, any score function $s(f_{\alpha,\beta}) = s(\alpha,\beta)$ of interest should be an increasing (or at least non-decreasing) function of β, and a decreasing (or at least non-increasing) function of α. In particular, we restrict our attention to score functions that are ‘separable’ in the sense that

$$s(\alpha,\beta) = a(\alpha)b(\beta), \tag{3.16}$$

where we take a to be a positive, continuous, decreasing, invertible function of α for $\alpha > \ell$, and b to be a positive, continuous, non-decreasing function of β.

Remark 1. This general Pareto model of Eq. (3.16) includes several natural score functions of interest, in particular:

i) In the case of the restricted Pareto distributions with finite mean, we may take s as the expected value, and $s(\alpha,\beta) = \alpha\beta/(\alpha-1)$, with $a(\alpha) = \alpha/(\alpha-1)$ and $b(\beta) = \beta$.

ii) In the case of unrestricted Pareto distributions, various asymptotic considerations give rise to considering the score function $s(\alpha,\beta) = 1/\alpha$, i.e., the controller attempts to find the arm with minimal α. In this case, $a(\alpha) = 1/\alpha$ and $b(\beta) = 1$. This arises for instance in comparing the asymptotic tail distributions of arms, $\mathbb{P}(X \geqslant k)$ as $k \to \infty$, or the conditional restricted expected values, $\mathbb{E}[X|X \leq k]$ as $k \to \infty$.

iii) A third score function to consider is the median, defined over unrestricted Pareto distributions, with $s(\alpha, \beta) = \beta 2^{1/\alpha}$, taking $a(\alpha) = 2^{1/\alpha}$, $b(\beta) = \beta$.

Given the above special cases, it is convenient to take the assumption when operating over $\mathscr{F}_\ell$ that $a(\alpha) \to \infty$ as $\alpha \to \ell$.

For $f = f_{\alpha,\beta} \in \mathscr{F}_\ell$, and a sample of size t of i.i.d. samples under f, we take the estimator $\hat{f}_t = f_{\hat{\alpha}_t, \hat{\beta}_t}$ where

$$\begin{aligned} \hat{\beta}_t &= \min_{n=1,\ldots,t} X_n, \\ \hat{\alpha}_t &= \frac{t-1}{\sum_{n=1}^t \ln\left(\frac{X_n}{\hat{\beta}_t}\right)}. \end{aligned} \tag{3.17}$$

The following result characterizes the distributions of these estimators; the proof is given in Section 3.5:

Lemma 4 *With $\hat{\alpha}_t, \hat{\beta}_t$ as in Eq.* (3.17)*, $\hat{\alpha}_t$ and $\hat{\beta}_t$ are independent, with*

$$\begin{aligned} \frac{\alpha}{\hat{\alpha}_t}(t-1) &\sim Gamma(t-1, 1), \\ \frac{\hat{\beta}_t}{\beta} &\sim Pareto(\alpha t, 1). \end{aligned} \tag{3.18}$$

It is convenient to define the following functions, $L^+(\delta), L^-(\delta)$, as respectively the smallest and largest positive solutions to $L - \ln L - 1 = \delta$ for $\delta \geqslant 0$. In particular, $L^-(\delta)$ may be expressed in terms of the Lambert-W function, $L^-(\delta) = -W(e^{-1-\delta})$, taking $W(x)$ be the principal solution to $We^W = x$ for $x \in [-1/e, \infty)$. An important property will be that $L^{\pm}(\delta)$ is continuous as a function of δ, and $L^{\pm}(\delta) \to 1$ as $\delta \to 0$.

Given the above, we may define the following policy as a specific instance of policy π^* under this model:

Policy π^* : UCB-PARETO(ℓ)

i) For $n = 1, 2, \ldots 3N$, pull each arm 3 times to construct initial estimators, and

ii) for $n \geqslant 3N$, pull arm $\pi^*(n+1) = \text{argmax}_i u_i\left(n, T^i_{\pi^*}(n)\right)$ breaking ties uniformly at random, where

$$u_i(n,t) = \begin{cases} \infty & \text{if } \hat{\alpha}^i_t L^-\left(\frac{\ln n}{t-2}\right) \leq \ell, \\ b\left(\hat{\beta}^i_t\right) a\left(\hat{\alpha}^i_t L^-\left(\frac{\ln n}{t-2}\right)\right) & \text{else.} \end{cases} \tag{3.19}$$

Theorem 6 *Policy π^*-UCB-PARETO(ℓ), as defined above is asymptotically optimal. In particular, for any choice of $\underline{f} \in \mathscr{F}^N_\ell$, with $f_i = f_{\alpha_i,\beta_i}$, with $s^* = \max_i s(\alpha_i,\beta_i) = \max_i a(\alpha_i)b(\beta_i)$, for each sub-optimal arm i the following holds:*

$$\lim_n \frac{\mathbb{E}_{\underline{f}}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\frac{1}{\alpha_i}a^{-1}\left(\frac{s^*}{b(\beta_i)}\right) - \ln\left(\frac{1}{\alpha_i}a^{-1}\left(\frac{s^*}{b(\beta_i)}\right)\right) - 1}. \tag{3.20}$$

Proof. It suffices to verify **Conditions 1-3**, and the continuity of s with respect to $\mathbf{I}$. To begin, it can be shown that

$$\mathbf{I}(f_{\alpha,\beta}, f_{\tilde{\alpha},\tilde{\beta}}) = \begin{cases} \frac{\tilde{\alpha}}{\alpha} - \ln\left(\frac{\tilde{\alpha}}{\alpha}\right) - 1 + \tilde{\alpha}\ln\left(\frac{\beta}{\tilde{\beta}}\right) & \text{if } \tilde{\beta} \leq \beta \\ \infty & \text{else,} \end{cases} \tag{3.21}$$

$$\mathbb{K}_{f_{\alpha,\beta}}(\rho) = \begin{cases} \frac{1}{\alpha}a^{-1}\left(\frac{\rho}{b(\beta)}\right) - \ln\left(\frac{1}{\alpha}a^{-1}\left(\frac{\rho}{b(\beta)}\right)\right) - 1 & \text{if } \rho > s(\alpha,\beta) \\ 0 & \text{else.} \end{cases} \tag{3.22}$$

The verification is simply computation and somewhat tedious, and is therefore relegated to Section 3.5. There are various choice of similarity measure ν that might be utilized to verify the relevant conditions; we take in this case the choice of $\nu = \mathbf{I}$.

3.3.2 Uniform Arms with (Semi)-Arbitrary Support

In this section, we consider an arm model that demonstrates the necessity of the general form of **Condition 3**. In particular, consider the set of distributions that are uniform over finite disjoint unions of closed sub-intervals of $[0,1]$, i.e.,

$$\mathscr{F} = \left\{ f_S = \mathbb{1}\left\{x \in S\right\}/|S| : S = \bigcup_{i=1}^{k}\left[a_i, b_i\right], 0 \leq a_1 < b_1 < \ldots < a_k < b_k < 1, k < \infty \right\}. \tag{3.23}$$

For S as in the above, we define $|S| = \sum_{i=1}^{k}(b_i - a_i)$ as the measure of S. We take as the score functional for this model $s(f_S) = |S|$, the measure of the support of the density $f_S \in \mathscr{F}$. To ensure that all relevant infima are defined, it is convenient to remove the complete interval $[0,1]$ from consideration, so we take $\mathscr{F}' = \mathscr{F} \setminus \{f_{[0,1]}\}$.

Given t i.i.d. samples from $f_S \in \mathscr{F}'$, we construct an estimator $\hat{f}_t$ of the form $\hat{f}_t = f_{\hat{S}_t}$, where $\hat{S}_t$ is an estimator of S constructed in the following way: Let d_k be a positive, integer valued, non-decreasing function that is unbounded and sub-linear in k. Consider a partition of $[0,1]$ into a sequence of intervals of width $\varepsilon_t = 1/d_t$. The estimator $\hat{S}_t$ is then taken to be the union of partition intervals that contain at least one sample of the t samples.

Given the above, we may define the following policy as a specific instance of policy π^* under this model: Defining $\tilde{d}(t) = d_t + 1$, let $n_0 = \min\{n : n > \tilde{d}(n)\}$.

Policy π^*: UCB-COVERAGE

i) For $n = 1, 2, \ldots n_0 \times N$, pull each arm n_0 times, and

ii) for $n \geqslant n_0 \times N$, pull arm $\pi^*(n+1) = \text{argmax}_i u_i\left(n, T^i_{\pi^*}(n)\right)$ breaking ties uniformly at random, where

$$u_i(n,t) = \max\left(|\hat{S}_t| n^{\frac{1}{t - \tilde{d}(t)}}, 1\right) \tag{3.24}$$

Theorem 7 *Policy π^*-UCB-COVERAGE as defined above is asymptotically optimal. In particular, for any choice of $\underline{f} \in (\mathscr{F}')^N$, with $f_i = f_{S_i}$, with $s^* = \max_i s(f_{S_i}) = \max_i |S_i|$, for each sub-optimal arm i the following holds:*

$$\lim_n \frac{\mathbb{E}_{\underline{f}}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\ln s^* - \ln|S_i|}. \tag{3.25}$$

Proof. It suffices to verify **Conditions 1-3** and the continuity of s relative to $\mathbf{I}$ for the indicated Uniform model. To begin, it can be shown that

$$\mathbf{I}(f_S, f_T) = \begin{cases} \ln|T| - \ln|S| & \text{if } S \subset T \\ \infty & \text{else} \end{cases}. \tag{3.26}$$

It follows easily from this that s is continuous under $\mathbf{I}$. Additionally, noting we are only interested in $\rho \leq 1$,

$$\mathbb{K}_{f_S}(\rho) = \begin{cases} \ln\rho - \ln|S| & \text{if } \rho > |S| \\ 0 & \text{else} \end{cases}. \tag{3.27}$$

In this case, we take as choice of similarity measure $\nu(f_S, f_T) = ||S| - |T||$. Given the form of $\mathbb{K}_{f_S}$ above, **Condition 1** is easily verified. The remainder of the verification, **Conditions 2 & 3**, is largely computation, though the verification of these conditions is of some interest in its dependence on the underlying distribution of $\hat{S}_t$ as an estimator. The full details are given in Section 3.5.

3.3.3 Uniform Arms of Interval Support

In this section, the uniform distributions are taken to be over single intervals, with finite but otherwise unconstrained bounds. This restriction to single intervals is necessary to ensure that the score functionals considered will be continuous with respect to **I**. We take $\mathscr{F}$ as the family of Uniform distributions with interval support:

$$\mathscr{F} = \{f_{a,b}(x) = \mathbb{1}\{x \in [a,b]\}/(b-a) : -\infty < a < b < \infty\}. \tag{3.28}$$

Taking X as distributed according to $f_{a,b} \in \mathscr{F}$, e.g., $X \sim \text{Unif}[a,b]$, X is distributed over $[a,b]$, with $\mathbb{E}[X] = (a+b)/2$. As this is a well defined function over all of $\mathscr{F}$, it makes for a reasonable (and traditional) score functional. However, we are aiming for greater generality. Taking the controller's goal as achieving large rewards from the pulled arms, any score functional $s(f_{a,b}) = s(a,b)$ of interest should be an increasing function of a, and an increasing function of b. We additionally take s to be continuous in a and b. Note, this is satisfied taking s as the expected value, $s_\mu(a,b) = (a+b)/2$.

For $f = f_{a,b} \in \mathscr{F}$, and t many i.i.d. samples under f, we take the estimator $\hat{f}_t = f_{\hat{a}_t,\hat{b}_t} \in \mathscr{F}$, where

$$\begin{aligned} \hat{a}_t &= \min_{n=1,\ldots,t} X_n, \\ \hat{b}_t &= \max_{n=1,\ldots,t} X_n. \end{aligned} \tag{3.29}$$

Given the above, we may define the following policy as a specific instance of policy π^* under this model:

Policy π^* : UCB-UNIFORM

i) For $n = 1, 2, \ldots 3N$, pull each arm 3 times, and

ii) for $n \geqslant 3N$, pull arm$\pi^*(n+1) = \text{argmax}_i u_i\left(n, T^i_{\pi^*}(n)\right)$ breaking ties uniformly at random, where

$$u_i(n,t) = s(\hat{a}^i_t, \hat{a}^i_t + n^{\frac{1}{t-2}}(\hat{b}^i_t - \hat{a}^i_t)). \tag{3.30}$$

Theorem 8 *For general s as outlined above, policy π^*-UNIFORM as defined above is asymptotically optimal. In particular, for any choice of $\underline{f} \in \mathscr{F}^N$, with $f_i = f_{a_i,b_i}$, with $s^* = \max_i s(a_i, b_i)$, for each sub-optimal arm i the following holds:*

$$\lim_n \frac{\mathbb{E}_{\underline{f}}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\min_{b_i \leq b}\left\{\ln\left(b - a_i\right) : s(a_i, b) \geqslant s^*\right\} - \ln(b_i - a_i)}. \tag{3.31}$$

Taking the particular choice of $s_\mu(a,b) = (a+b)/2$, this yields for all sub-optimal i,

$$\lim_n \frac{\mathbb{E}_{\underline{f}}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\ln\left(\frac{2s^* - 2a_i}{b_i - a_i}\right)}. \tag{3.32}$$

This uniform case, in terms of regret minimization, appears in more detail with additional finite horizon regret bounds in [14].

Proof. It suffices to verify **Conditions 1-3**, and the continuity of s under $\mathbf{I}$. It can be shown that

$$\mathbf{I}(f_{a,b}, f_{\tilde{a},\tilde{b}}) = \begin{cases} \ln\left(\frac{\tilde{b}-\tilde{a}}{b-a}\right) & \text{if } \tilde{a} \leq a, b \leq \tilde{b} \\ \infty & \text{else,} \end{cases} \tag{3.33}$$

$$\mathbb{K}_{f_{a,b}}(\rho) = \min_{b \leq \tilde{b}}\left\{\ln\left(\tilde{b} - a\right) : s(a, \tilde{b}) \geqslant \rho\right\} - \ln(b-a).$$

The remainder of the verification is simply computation and somewhat tedious, and is therefore relegated to Section 3.5. There are various choice of similarity measure ν that might be utilized to verify the relevant conditions; we take in this case the choice of $\nu = \mathbf{I}$.

3.3.4 Normal Arms of Unknown Mean, Unknown Variance

In this section, we consider the results of a given arm to be i.i.d., normally distributed, with unknown mean and unknown variance. That is, we take

$$\mathscr{F} = \left\{ f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty \right\}. \tag{3.34}$$

Additionally, we take the score functional of interest to be the mean or expected value, $s(f_{\mu,\sigma}) = s(\mu,\sigma) = \mu$. Given t i.i.d. samples from $f_{\mu,\sigma}$, we take as an estimator $\hat{f}_t = f_{\hat{\mu}_t,\hat{\sigma}^2(t)}$, where

$$\begin{aligned} \hat{\mu}_t &= \bar{X}_t = \frac{1}{t}\sum_{k=1}^{t} X_k, \\ \hat{\sigma}^2(t) &= \frac{1}{t-1}\sum_{k=1}^{t}(X_k - \hat{\mu}_t)^2. \end{aligned} \tag{3.35}$$

The problem of normal arms with unknown mean but *known* variances, i.e., arm i is known to have variance σ_i^2, was solved in [41], with regards to regret minimization, where a policy based on the index

$$u_i^{\mathrm{KR}}(n,t) = \bar{X}_t^i + \sigma_i\sqrt{\frac{2\ln n}{t}} \tag{3.36}$$

was shown to be asymptotically optimal. It additionally follows from that proof (and can be shown from Theorem 5), that such a policy also is asymptotically optimal with respect to mistake minimization, i.e., achieves the lower bound given in Eq. (2.26). The problem of regret minimization with respect to unknown variances remained open for some time, however, [9] conjecturing that an index policy based on

$$u_i^{\mathrm{BK}}(n,t) = \bar{X}_t^i + \hat{\sigma}_i(t)\sqrt{n^{\frac{2}{t}} - 1}, \tag{3.37}$$

would be asymptotically optimal. However, it was shown in [12] that such a policy is *not* asymptotically optimal, incurring regret that grew as a power of n, rather than logarithmically. It was additionally shown in [12] that a policy based on an index of the form

$$u_i^{\mathrm{CHK}}(n,t) = \bar{X}_t^i + \hat{\sigma}_i(t)\sqrt{n^{\frac{2}{t-2}} - 1}, \tag{3.38}$$

was in fact asymptotically optimal, achieving a minimal logarithmic growth of regret. Cowan et al. in [12] present a more detailed analysis of this policy. Applying the policy to mistake minimization, and noting however that u_i^{CHK} is precisely of the form given in Eq. (3.5), we produce a concise proof of asymptotic optimality here, as an application of Theorem 5.

Policy $\pi_{\mathbf{CHK}}$: UCB-NORMAL

i) For $n = 1, 2, \ldots 3N$, pull each arm 3 times, and

ii) for $n \geqslant 3N$, pull arm $\pi_{\mathrm{CHK}}(n+1) = \mathrm{argmax}_i u_i^{\mathrm{CHK}}\left(n, T^i_{\pi_{\mathrm{CHK}}}(n)\right)$ breaking ties uniformly at random.

Theorem 9 *For $s(f_{\mu,\sigma}) = \mu$ in the above model, policy π_{CHK} as defined above is asymptotically optimal. In particular, for any choice of $\underline{f} \in \mathscr{F}^N$, with $f_i = f_{\mu_i,\sigma_i}$, with $\mu^* = \max_i \mu_i$, for each sub-optimal arm i the following holds:*

$$\lim_n \frac{\mathbb{E}_{\underline{f}}\left[T^i_{\pi_{CHK}}(n)\right]}{\ln n} = \frac{2}{\ln\left(1 + \frac{(\mu^* - \mu_i)^2}{\sigma_i^2}\right)}. \tag{3.39}$$

Proof. It suffices to verify **Conditions 1-3**, and the continuity of s under $\mathbf{I}$. For this model, it can be shown that

$$\begin{aligned} \mathbf{I}(f,g) &= \frac{(\mu_f - \mu_g)^2}{2\sigma_g^2} + \frac{1}{2}\left(\frac{\sigma_f^2}{\sigma_g^2} - \ln\left(\frac{\sigma_f^2}{\sigma_g^2}\right) - 1\right), \\ \mathbb{K}_{f_{\mu,\sigma}}(\rho) &= \begin{cases} \frac{1}{2}\ln\left(1 + \frac{(\rho-\mu)^2}{\sigma^2}\right) & \text{if } \rho > \mu \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{3.40}$$

Let $L^-(\delta)$ and $L^+(\delta)$ be the smallest and largest positive solutions to $L - \ln L - 1 = \delta$, respectively. It follows then that if $\mathbf{I}(f,g) < \delta$,

$$\begin{aligned} \frac{(\mu_f - \mu_g)^2}{2\sigma_g^2} &< \delta, \\ \frac{1}{2}\left(\frac{\sigma_f^2}{\sigma_g^2} - \ln\left(\frac{\sigma_f^2}{\sigma_g^2}\right) - 1\right) &< \delta. \end{aligned} \tag{3.41}$$

From this, we have that

$$\begin{aligned} \sigma_f^2/L^+(2\delta) < \sigma_g^2 &< \sigma_f^2/L^-(2\delta), \\ |\mu_f - \mu_g| < \sigma_g\sqrt{2\delta} &< \sigma_f\sqrt{2\delta/L^{-1}(2\delta)}. \end{aligned} \tag{3.42}$$

Since $L^{\pm}(2\delta) \to 1$ and $\delta/L^-(2\delta) \to 0$ as $\delta \to 0$, the above implies that any functional of normal densities that is continuous with respect to the parameters will be continuous with respect to f

under $\mathbf{I}$. This immediately verifies the continuity of $s(f_{\mu,\sigma}) = \mu$ under $\mathbf{I}$. Additionally, $\mathbb{K}_{f_{\mu,\sigma}}$ is continuous with respect to ρ by inspection, and continuous with respect to f under $\mathbf{I}$ as well, by the previous remarks. Taking $\nu = \mathbf{I}$, this verifies **Condition 1**. To verify **Condition 2**, note that from similar analysis to the above,

$$\begin{aligned} \mathbb{P}_f\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) \leq \mathbb{P}_f\left(\frac{(\hat{\mu}_t - \mu)^2}{\sigma^2} > 2\delta\right) \\ + \mathbb{P}_f\left(\frac{\hat{\sigma}^2(t)}{\sigma^2} < L^-(2\delta)\right) + \mathbb{P}_f\left(\frac{\hat{\sigma}^2(t)}{\sigma^2} > L^+(2\delta)\right). \end{aligned} \tag{3.43}$$

Recalling that in the case of normals, $\hat{\mu}_t \sim \mu + Z\sigma/\sqrt{t}$, and $\hat{\sigma}^2(t) \sim \sigma^2 U_{t-1}/(t-1)$ where Z is a standard normal random variable, and U_{t-1} is an independent, χ^2_{t-1} random variable,

$$\begin{aligned} \mathbb{P}_f\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) &\leq \mathbb{P}_f\left(Z^2 > 2\delta t\right) \\ &\quad + \mathbb{P}_f\left(\frac{U_{t-1}}{t-1} < L^-(2\delta)\right) + \mathbb{P}_f\left(\frac{U_{t-1}}{t-1} > L^+(2\delta)\right) \\ &\leq e^{-\Omega(t)} + e^{-\Omega(t)} + e^{-\Omega(t)} \leq e^{-\Omega(t)} \end{aligned} \tag{3.44}$$

The exponential concentration results for each of the three terms are proven as Lemma 7 in Section 3.5, but are simply standard Chernoff-type bounds. This verifies **Condition 2**.

It remains to verify **Condition 3**, perhaps the most interesting of the three. Note that

$$\begin{aligned} \mathbb{P}_f\left(\delta < \mathbb{K}_{\hat{f}_t}(\rho)\right) &= \mathbb{P}_f\left(\delta < \frac{1}{2}\ln\left(1 + \frac{(\rho - \hat{\mu}_t)^2}{\hat{\sigma}_t^2}\right) \text{ and } \rho > \hat{\mu}_t\right) \\ &= \mathbb{P}_f\left(\hat{\sigma}_t\sqrt{e^{2\delta} - 1} < |\rho - \hat{\mu}_t| \text{ and } \rho > \hat{\mu}_t\right) \\ &= \mathbb{P}_f\left(\hat{\mu}_t + \hat{\sigma}_t\sqrt{e^{2\delta} - 1} < \rho\right). \end{aligned} \tag{3.45}$$

Hence,

$$
\begin{aligned}
\mathbb{P}_f\left(\delta < \mathbb{K}_{\hat{f}_t}(\mu - \varepsilon)\right) &= \mathbb{P}_f\left(\hat{\mu}_t + \hat{\sigma}_t\sqrt{e^{2\delta} - 1} < \mu - \varepsilon\right) \\
&= \mathbb{P}_f\left(Z\sigma/\sqrt{t} + \hat{\sigma}_t\sqrt{e^{2\delta} - 1} < -\varepsilon\right) \\
&= \mathbb{P}_f\left(\frac{\varepsilon}{\sigma}\sqrt{t} + \frac{\hat{\sigma}_t}{\sigma}\sqrt{t}\sqrt{e^{2\delta} - 1} < Z\right) \\
&\leq \mathbb{E}_f\left[e^{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\sqrt{t} + \frac{\hat{\sigma}_t}{\sigma}\sqrt{t}\sqrt{e^{2\delta}-1}\right)^2}\right] \\
&\leq e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}t}\mathbb{E}_f\left[e^{-\frac{1}{2}\frac{\hat{\sigma}_t^2}{\sigma^2}t\left(e^{2\delta}-1\right)}\right] \\
&= e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}t}\mathbb{E}\left[e^{-\frac{1}{2}U_{t-1}\frac{t}{t-1}\left(e^{2\delta}-1\right)}\right] \\
&= e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}t}\left(\frac{t-1}{e^{2\delta}t - 1}\right)^{\frac{t-1}{2}} \\
&\leq e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2}t}e^{-\delta(t-1)}.
\end{aligned}
\tag{3.46}
$$

The last step follows, taking $\delta > 0$. This verifies **Condition 3**, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$. Equation (3.39) follows from Eq. (3.40), the definition of $\mathbb{K}_f(\rho)$ for this model.

3.4 Normal Arms and a Joint UCB Sampling Policy

In the previous sections, the focus has been on contexts in which the arms shared no known relationship or structure. This proved an advantage in the analysis of the UCB Index policy π^*, as the index of one arm could be considered without regards to the data collected on the other arms, and thus each arm could effectively be considered in isolation. In this section, we consider generalizing the results of the previous sections to construct a UCB Index policy for arms with known structure, i.e., taking the set of plausible arm hypotheses to be $\mathscr{F}_P^N$, with non-trivial P. While we do not prove a general optimality result as done for the trivial-P case, we demonstrate asymptotic optimality for an interesting case involving normal arms.

Ignoring the $o(t)$ term for now - it will be dealt with shortly - the constraint on the optimization problem defined by $u_i(n,t)$ in Eq. (3.5) is $t\,\mathbf{I}(\hat{F}_t^i, G) \leq \ln n$. Interpreting $\mathbf{I}$ as the *limiting average log likelihood ratio*, this condition can be approximated as *the likelihood ratio between $\hat{F}_t^i$ and G*

is at most n. Taking $\hat{F}^i_t$ as the maximum likelihood estimator for F_i given the collected samples, this offers the (approximate) interpretation of $u_i(n,t)$ as *given the current data, the largest score of any hypothesis* $G \in \mathscr{F}$ *no less likely than a fraction of the maximal likelihood of any hypothesis for i*. This suggests an immediate extension to the structured arm case: given $\underline{t} = (t_1, \dots, t_N)$ pulls from each arm, let $\underline{\hat{F}}(\underline{t}) \in \mathscr{F}^N_P$ be the maximum likelihood estimator for $\underline{F} \in \mathscr{F}^N_P$, then take the index to be *given the current data, the largest i-score for any arm hypotheses* $\underline{G} \in \mathscr{F}^N_P$ *no less likely than a fraction of the likelihood of* $\underline{\hat{F}}(\underline{t})$. Using $\mathbf{I}$ to approximate the average log likelihood ratios again, and reintroducing the sub-linear terms, this leads to the following 'joint' index:

$$v_i(n,\underline{t}) = \sup_{\underline{G} \in \mathscr{F}^N_P} \left\{ s(G_i) : \sum_{j=1}^{N} (t_j - o(t_j)) \mathbf{I}\left(\underline{\hat{F}}(\underline{t})_j, G_j\right) \leq \ln n \right\}. \tag{3.47}$$

Remark 2. In the case that P is trivial, the above recovers the previous UCB index for the unrelated case, $v_i(n,\underline{t}) = u_i(n,t_i)$, as we may take $G_j = \underline{\hat{F}}(\underline{t})_j$ for each $j \neq i$.

Remark 3. Additionally, we include the sub-linear $o(t_j)$ terms, due to the proven necessity in the unrelated arm case (Section 3.3). While $t\ \mathbf{I}(\hat{F}_t, G)$ more directly approximates the log likelihood ratio, the inclusion of the $-o(t)$ term seems necessary as an 'unbiasing' effect, to ensure the correct mistake rates are achieved. The above is the natural generalization to the related arm, non-trivial P case.

Analysis of a policy based on the above index v_i is complicated by the fact that the index for arm i depends explicitly on the data from the other arms. While we expect an index policy based on v_i to generally perform quite well - perhaps even optimally, given the appropriate choices of the $o(t_i)$ terms, we present no general optimality result, but consider the following special case:

We again restrict to the i.i.d. case, and in doing so may identify $\mathscr{F}$ with the underlying density functions for a single pull. Consider the case where the arms are known to be i.i.d. normals, with finite means and unknown variance, and additionally the variances are known to be equal between arms. That is, we take $\mathscr{F}$ as the set of i.i.d. normal densities with finite mean and variance, and take $\underline{f}$ to satisfy P as $\text{Var}_{f_1}(X) = \text{Var}_{f_2}(X) = \ldots = \text{Var}_{f_N}(X)$. In this context, we take the score function to be the mean, $s(f) = \mu(f) = \mathbb{E}_f[X]$.

Note, in this case, we have already shown as an application of Theorem 3, and Eq. (2.27), that for any uniformly fast policy π we have the following bound, taking $\underline{f} \in \mathscr{F}^N_P$, with $f_i = f_{\mu_i,\sigma}$

and $\mu^* = \max_i \mu_i$,

$$\liminf_n \frac{M_\pi^{\underline{f}}(n)}{\ln n} \geqslant \sum_{i:\mu_i \neq \mu^*} \frac{2\sigma^2}{(\mu^* - \mu_i)^2}. \tag{3.48}$$

We will show that an index policy in this case based on v_i is in fact asymptotically optimal, achieving this indicated bound. In this framework, we construct the following estimators, given $\underline{t} = (t_1, \ldots, t_N)$ samples from each arm:

$$\begin{aligned} \hat{\mu}^i_{t_i} &= \frac{1}{t_i} \sum_{k=1}^{t_i} X^i_k \\ \hat{\sigma}^2_i(t_i) &= \frac{1}{t_i} \sum_{k=1}^{t_i} \left(X^i_k - \hat{\mu}^i_{t_i}\right)^2 \\ \hat{\sigma}^2(\underline{t}) &= \frac{\sum_{i=1}^N t_i \hat{\sigma}^2_i(t_i)}{\sum_{i=1}^N (t_i - 1)}. \end{aligned} \tag{3.49}$$

That is, for each arm i we estimate the mean and variance with the sample mean and sample variance for that arm, and we additionally construct a pooled estimate of the common variance from the estimators for each arm. We then define the following policy, based on the specific instance of v_i above in this case, taking each for each i the $o(t_i) = 1$:

Policy π^*: UCB-NORMAL-COMMON-VARIANCE

i) For $n = 1, 2, \ldots 2N$, pull each arm 2 times, and

ii) for $n \geqslant 2N$, pull arm $\pi^*(n+1) = \text{argmax}_i v_i\left(n, (T^1_{\pi^*}(n), \ldots, T^N_{\pi^*}(n))\right)$ breaking ties uniformly at random, where

$$v_i(n, \underline{t}) = \hat{\mu}^i_{t_i} + \frac{\hat{\sigma}(\underline{t})}{\sqrt{t_i - 1}} \sqrt{(n-N)\left(n^{\frac{2}{n-N}} - 1\right)}. \tag{3.50}$$

Theorem 10 *In the context outlined above, policy π^*-NORMAL-COMMON-VARIANCE as defined above is asymptotically optimal. In particular, for any choice of $\underline{f} \in \mathscr{F}_P^N$, with $f_i = f_{\mu_i,\sigma}$, with $\mu^* = \max_i \mu_i$, the following holds:*

$$\lim_n \frac{M_{\pi^*}^{\underline{f}}(n)}{\ln n} = \sum_{i:\mu_i \neq \mu^*} \frac{2\sigma^2}{(\mu^* - \mu_i)^2}. \tag{3.51}$$

The proof is given in Section 3.5. It is similar to the proof of Theorem 4, but differs in significant ways due to the index depending on the pooled data of all arms, rather than a single arm alone.

Remark 4. This is a remarkable result: this optimal asymptotic mistake rate for the indicated policy is *identical* to the optimal asymptotic mistake rate in the case of normal arms with unknown means and known variances (Eq. (2.26)). That is, correctly utilized (through the form of the index function), the knowledge that the variances are equal is as useful in asymptotically minimizing the mistakes as knowing exactly what that variance is. Any cost due to the additional uncertainty of not knowing the true variance is restricted to sub-logarithmic order mistakes. This is a delightful result.

3.5 Proofs

Proof. [of Theorem 4.] We recall the definition of $\mathbb{K}_F(\rho)$, and introduce a companion function, $C_F(\delta)$:

$$\begin{aligned} \mathbb{K}_F(\rho) &= \inf_{G\in\mathscr{F}} \left\{ \mathbf{I}(F,G) : s(G) > \rho \right\}, \\ C_F(\delta) &= \sup_{G\in\mathscr{F}} \left\{ s(G) : \mathbf{I}(F,G) < \delta \right\}. \end{aligned} \tag{3.52}$$

Thinking of $\mathbb{K}_F(\rho)$ as the minimal distance (relative to $\mathbf{I}$) from F to a law better than ρ, we may consider $C_F(\delta)$ to be the best score achieved within distance δ of F. Note, we have the following relationship: $u_i(n,t) = C_{\hat{F}^i_t}(\ln n/(t-\tilde{d}(t)))$. Note as well, $\mathbb{K}_F(\rho)$ is an increasing function with ρ, and $\mathbb{K}_F(C_F(\delta)) \leq \delta$.

Fix $\underline{F} = (F_1,\dots,F_N) \in \mathscr{F}^N$, with $i \notin O(\underline{F})$ as sub-optimal arm relative to $\underline{F}$, and $i^* \in O(\underline{F})$ an optimal arm. For convenience, we take $s^* = s^*(\underline{F})$. Let ε, δ be feasible as in the statement of the Theorem. We define the following functions, for $n \geqslant n_0 N$:

$$\begin{aligned} n^i_1(n,\varepsilon,\delta) &= \sum_{t=n_0N}^{n} \mathbb{1}\left\{ \pi^*(t+1) = i, u_i(t,T^i_{\pi^*}(t)) \geqslant s^* - \varepsilon, \nu(\hat{F}^i_{T^i_{\pi^*}(t)},F_i) \leq \delta \right\} \\ n^i_2(n,\varepsilon,\delta) &= \sum_{t=n_0N}^{n} \mathbb{1}\left\{ \pi^*(t+1) = i, u_i(t,T^i_{\pi^*}(t)) \geqslant s^* - \varepsilon, \nu(\hat{F}^i_{T^i_{\pi^*}(t)},F_i) > \delta \right\} \\ n^i_3(n,\varepsilon) &= \sum_{t=n_0N}^{n} \mathbb{1}\left\{ \pi^*(t+1) = i, u_i(t,T^i_{\pi^*}(t)) < s^* - \varepsilon \right\}. \end{aligned} \tag{3.53}$$

Note the relation that $T^i_{\pi^*}(n+1) = n_0 + n^i_1(n,\varepsilon,\delta) + n^i_2(n,\varepsilon,\delta) + n^i_3(n,\varepsilon)$.

We have the following relation:

$$
\begin{aligned}
&\left\{u_i(t,k) \geqslant s^* - \varepsilon, \nu(\hat{F}^i_k, F_i) \leq \delta\right\} \\
&= \left\{C_{\hat{F}^i_k}(\ln t/(k - \tilde{d}(k))) \geqslant s^* - \varepsilon, \nu(\hat{F}^i_k, F_i) \leq \delta\right\} \\
&= \left\{\mathbb{K}_{\hat{F}^i_k}(C_{\hat{F}^i_k}(\ln t/(k - \tilde{d}(k)))) \geqslant \mathbb{K}_{\hat{F}^i_k}(s^* - \varepsilon), \nu(\hat{F}^i_k, F_i) \leq \delta\right\} \\
&\subset \left\{\ln t/(k - \tilde{d}(k)) \geqslant \mathbb{K}_{\hat{F}^i_k}(s^* - \varepsilon), \nu(\hat{F}^i_k, F_i) \leq \delta\right\} \\
&\subset \left\{\ln t/(k - \tilde{d}(k)) \geqslant \inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}\right\} \\
&= \left\{\ln t/ \inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\} + \tilde{d}(k) \geqslant k\right\}
\end{aligned}
\tag{3.54}
$$

This gives us the following bounds:

$$
\begin{aligned}
&n^i_1(n, \varepsilon, \delta) \\
&\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{\pi^*(t+1) = i, \frac{\ln t}{\inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}} + \tilde{d}(T^i_{\pi^*}(t)) \geqslant T^i_{\pi^*}(t)\right\} \\
&\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{\pi^*(t+1) = i, \frac{\ln n}{\inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}} + \tilde{d}(T^i_{\pi^*}(t)) \geqslant T^i_{\pi^*}(t)\right\} \\
&\leq \sum_{t=0}^{n} \mathbb{1}\left\{\pi^*(t+1) = i, \frac{\ln n}{\inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}} + \tilde{d}(T^i_{\pi^*}(t)) \geqslant T^i_{\pi^*}(t)\right\} \\
&\leq \max\left\{T : T - \tilde{d}(T) \leq \frac{\ln n}{\inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}}\right\} + 1.
\end{aligned}
\tag{3.55}
$$

The last bounds in the above hold with the following reasoning: Viewing $T^i_{\pi^*}(t)$ as the sum of $\mathbb{1}\{\pi^*(t) = i\}$ terms, the added conditioning in the above indicators restrict how many terms of the above sum can be non-zero, hence how large $T^i_{\pi^*}(t)$ can be for any t. The inclusion of the $+1$ term in the last step accounts for the $\pi^*(n+1)$ term present in the above sum, not present in the sum for $T^i_{\pi^*}(n)$. Note, this bound holds almost surely, independent of outcomes. Further then, taking $\tilde{d}$ as positive and increasing, for any positive C, we have the relation that $\max\{T : T - \tilde{d}(T) \leq C\} \leq C + O(\tilde{d}(C))$. Hence, since $\tilde{d}$ is taken to be sub-linear,

$$
n^i_1(n, \varepsilon, \delta) \leq \frac{\ln n}{\inf_{G \in \mathscr{F}} \{\mathbb{K}_G(s^* - \varepsilon) : \nu(G, F_i) \leq \delta\}} + o(\ln n).
\tag{3.56}
$$

To bound the n_2^i term, observe the following:

$$\begin{aligned}
n_2^i(n,\varepsilon,\delta) &\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{\pi^*(t+1)=i, \nu(\hat{F}^i_{T^i_{\pi^*}(t)},F_i) > \delta\right\} \\
&= \sum_{t=n_0N}^{n}\sum_{k=n_0}^{t} \mathbb{1}\left\{\pi^*(t+1)=i, \nu(\hat{F}^i_k,F_i) > \delta, T^i_{\pi^*}(t)=k\right\} \\
&= \sum_{t=n_0N}^{n}\sum_{k=n_0}^{t} \mathbb{1}\left\{\pi^*(t+1)=i, T^i_{\pi^*}(t)=k\right\}\mathbb{1}\left\{\nu(\hat{F}^i_k,F_i) > \delta\right\} \\
&\leq \sum_{k=n_0}^{n} \mathbb{1}\left\{\nu(\hat{F}^i_k,F_i) > \delta\right\}\sum_{t=k}^{n} \mathbb{1}\left\{\pi^*(t+1)=i, T^i_{\pi^*}(t)=k\right\} \\
&\leq \sum_{k=n_0}^{n} \mathbb{1}\left\{\nu(\hat{F}^i_k,F_i) > \delta\right\}.
\end{aligned} \tag{3.57}$$

The last step of the above follows, as for any k, $\pi^*(t+1)=i, T^i_{\pi^*}(t)=k$ can be true for at most one t.

To bound the n_3^i term, note that by the structure of the policy, if $\pi^*(t+1)=i$, then $u_i(t,T^i_{\pi^*}(t)) = \max_j u_j(t,T^j_{\pi^*}(t))$. Hence, if i^* is an optimal arm, $\pi^*(t+1)=i$, and $u_i(t,T^i_{\pi^*}(t)) < s^*-\varepsilon$, it must also be that $u_{i^*}(t,T^{i^*}_{\pi^*}(t)) < s^*-\varepsilon$. Hence we have the following bound:

$$\begin{aligned}
n_3^i(n,\varepsilon) &\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{\pi^*(t+1)=i, u_{i^*}(t,T^{i^*}_{\pi^*}(t)) < s^*-\varepsilon\right\} \\
&\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{u_{i^*}(t,T^{i^*}_{\pi^*}(t)) < s^*-\varepsilon\right\} \\
&\leq \sum_{t=n_0N}^{n} \mathbb{1}\left\{u_{i^*}(t,k) < s^*-\varepsilon \text{ for some } k=n_0,\ldots,t\ \right\} \\
&\leq \sum_{t=n_0N}^{n}\sum_{k=n_0}^{t} \mathbb{1}\left\{u_{i^*}(t,k) < s^*-\varepsilon\right\}.
\end{aligned} \tag{3.58}$$

Combining each of the above bounds, and observing that $T^i_{\pi^*}(n) \leq T^i_{\pi^*}(n+1)$, we have for $n \geqslant n_0N$:

$$\begin{aligned}
T^i_{\pi^*}(n) \leq &\frac{\ln n}{\inf_{G\in\mathscr{F}}\left\{\mathbb{K}_G(s^*-\varepsilon) : \nu(G,F_i) \leq \delta\right\}} + o(\ln n) \\
&+ \sum_{k=n_0}^{n} \mathbb{1}\left\{\nu(\hat{F}^i_k,F_i) > \delta\right\} \\
&+ \sum_{t=n_0N}^{n}\sum_{k=n_0}^{t} \mathbb{1}\left\{u_{i^*}(t,k) < s^*-\varepsilon\right\}.
\end{aligned} \tag{3.59}$$

Taking expectations completes the proof.

Proposition 3 *For* $\Delta > 0, \tilde{d}(k) = o(k), t > 1,$

$$\sum_{k=1}^{\infty} t^{-\Delta/\left(k-\tilde{d}(k)\right)} e^{-\Omega(k)} \leq O(1/\ln t). \tag{3.60}$$

Proof. Let $1 > p > 0$. We have

$$\begin{aligned}\sum_{k=1}^{\infty} t^{-\Delta/\left(k-\tilde{d}(k)\right)} e^{-\Omega(k)} &= \sum_{k=1}^{\lfloor \ln(t)^p \rfloor} t^{-\Delta/\left(k-\tilde{d}(k)\right)} e^{-\Omega(k)} + \sum_{k=\lceil \ln(t)^p \rceil}^{\infty} t^{-\Delta/\left(k-\tilde{d}(k)\right)} e^{-\Omega(k)} \\ &\leq \sum_{k=1}^{\lfloor \ln(t)^p \rfloor} t^{-\Delta/\left(k-\tilde{d}(k)\right)} + \sum_{k=\lceil \ln(t)^p \rceil}^{\infty} e^{-\Omega(k)} \\ &= \ln(t)^p e^{-\Omega(\ln(t)^{1-p})} + e^{-\Omega(\ln(t)^p)}.\end{aligned} \tag{3.61}$$

As the exponential function decays faster than any polynomial of its argument, we have from the above immediately that both of the above terms are $o((\ln t)^{-\alpha})$ for any $\alpha > 0$. Hence, taking $\alpha = 1$,

$$\sum_{k=1}^{\infty} t^{-\Delta/(k-\tilde{d}(k))} e^{-\Omega(k)} \leq O(1/\ln(t)). \tag{3.62}$$

Proof. [of Lemma 4.] To see the distribution of $\hat{\alpha}_n$, consider the event that $X_1 = \min_t X_t$. This can be generated in the following way, by first generating X_1 according to Pareto(α, β), then for each $j \neq 1$, generating each X_j independently as Pareto(α, β) conditioned on $X_j \geqslant X_1$, in which case $X_j \sim$ Pareto(α, X_1), by the self-similarity of the Pareto distribution. Using the standard fact that if $X \sim$ Pareto(α, β), then $\ln(X/\beta) \sim$ Exp(α), we have that

$$\sum_{t=1}^{n} \ln\left(\frac{X_t}{X_1}\right) \tag{3.63}$$

is distributed as the sum of $n-1$ many i.i.d. exponential random variables with parameter α, or Gamma$(n-1, \alpha)$. Note, this holds independent of the value of X_1. The same argument holds, taking any of the X_t as the minimum. Hence, independent of which X_t is the minimum, and independent of the value of that minimum (i.e., independent of $\hat{\beta}_n$, the above sum is distributed like Gamma$(n-1, \alpha) \sim$ Gamma$(n-1, 1)/\alpha$. This gives the above representation of $\hat{\alpha}_n$ and demonstrates the independence of $\hat{\alpha}_n$ and $\hat{\beta}_n$.

To see the distribution of $\hat{\beta}_n$, note that $\hat{\beta}_n \geqslant \beta$, and for $x \geqslant 1$,

$$\mathbb{P}(\hat{\beta}_n/\beta > x) = \mathbb{P}(\hat{\beta}_n > \beta x) = \prod_{t=1}^{n} \mathbb{P}(X_t > \beta x) = \left(\frac{\beta}{\beta x}\right)^{n\alpha} = \left(\frac{1}{x}\right)^{n\alpha}, \tag{3.64}$$

which shows that $\hat{\beta}_n/\beta \sim \text{Pareto}(\alpha n, 1)$.

Proof. [of Theorem 6.] Note that $\mathbf{I}(f_{\alpha,\beta}, f_{\tilde{\alpha},\tilde{\beta}}) < \delta$ implies that

$$\begin{gathered}
\tilde{\beta} \leq \beta \\
\frac{\tilde{\alpha}}{\alpha} - \ln\left(\frac{\tilde{\alpha}}{\alpha}\right) - 1 \leq \delta \\
\tilde{\alpha} \ln\left(\frac{\beta}{\tilde{\beta}}\right) \leq \delta.
\end{gathered} \tag{3.65}$$

The above gives us that $\alpha L^-(\delta) \leq \tilde{\alpha} \leq \alpha L^+(\delta)$ and $\beta e^{-\alpha\delta L^+(\delta)} \leq \tilde{\beta} \leq \beta$. Given that $\delta L^+(\delta) \to 0$ as $\delta \to 0$, these bounds and the continuity of a, b, give the continuity of s with respect to $\mathbf{I}$.

It is convenient to take as similarity measure on $\mathscr{F}_\ell$, $\nu = \mathbf{I}$. **Condition 1** is then easily verified, the continuity of $\mathbb{K}_f(\rho)$ with respect to ρ from the given formula for $\mathbb{K}$, and the continuity with respect to f under $\mathbf{I}$ from the previous bounds.

In verifying **Condition 2**, it is interesting to note that for $\ell > 0$, the estimator $\hat{f}_t = f_{\hat{\alpha}_t,\hat{\beta}_t}$ of $f = f_{\alpha,\beta}$ may not be in $\mathscr{F}_\ell$ even if f is, i.e., even if $\alpha > \ell$, there is no immediate guarantee that $\hat{\alpha}_t$ is. Hence, $\mathbf{I}(\hat{f}_t, f)$ may not be well defined over $\mathscr{F}_\ell$. However, this is not a serious issue as in the case that $\ell > 0$, we may view this as embedded in $\mathscr{F}_0$, which will contain $\hat{f}_t$, and hence allow us to compute $\mathbf{I}(\hat{f}_t, f)$. Hence, for $\delta > 0$, since $\hat{\beta}_t \geqslant \beta$,

$$\begin{aligned}
&\mathbb{P}_f\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) \\
&= \mathbb{P}_f\left(\frac{\alpha}{\hat{\alpha}_t} - \ln\left(\frac{\alpha}{\hat{\alpha}_t}\right) - 1 + \alpha \ln\left(\frac{\hat{\beta}_t}{\beta}\right) > \delta\right) \\
&\leq \mathbb{P}_f\left(\frac{\alpha}{\hat{\alpha}_t} - \ln\left(\frac{\alpha}{\hat{\alpha}_t}\right) - 1 > \frac{\delta}{2}\right) + \mathbb{P}_f\left(\alpha \ln\left(\frac{\hat{\beta}_t}{\beta}\right) > \frac{\delta}{2}\right) \\
&= \mathbb{P}_f\left(\frac{\alpha}{\hat{\alpha}_t} < L^-\left(\frac{\delta}{2}\right)\right) + \mathbb{P}_f\left(\frac{\alpha}{\hat{\alpha}_t} > L^+\left(\frac{\delta}{2}\right)\right) + \mathbb{P}_f\left(\frac{\hat{\beta}_t}{\beta} > e^{\frac{\delta}{2\alpha}}\right).
\end{aligned} \tag{3.66}$$

Recalling the characterizations of the distributions of $\hat{\alpha}_t$ and $\hat{\beta}_t$ (Lemma 4), letting $G_t \sim \text{Gamma}(t, 1)$,

$$\mathbb{P}_f\left(\mathbf{I}(\hat{f}_{t+1}, f) > \delta\right) \leq \mathbb{P}\left(G_t < (t)L^-(\delta/2)\right) + \mathbb{P}\left(G_t > tL^+(\delta/2)\right) + e^{-\frac{\delta}{2}(t+1)}. \tag{3.67}$$

Here we apply the following result, bounding the tails of the Gamma distributions:

Lemma 5 *Let $G_t \sim Gamma(t,1)$. For $0 < \gamma^- < 1 < \gamma^+ < \infty$, the following bounds hold:*

$$\begin{aligned}\mathbb{P}\left(G_t < t\gamma^-\right) &\le \left(\gamma^- e^{1-\gamma^-}\right)^t \\ \mathbb{P}\left(G_t > t\gamma^+\right) &\le \left(\gamma^+ e^{1-\gamma^+}\right)^t.\end{aligned} \tag{3.68}$$

These are standard Chernoff bounds, the proof given following this one. Applying them to the above, taking $\gamma^\pm = L^\pm(\delta/2)$, note that $\gamma^\pm e^{1-\gamma^\pm} = e^{-\delta/2}$. Hence,

$$\mathbb{P}\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) \le 2e^{-\frac{\delta}{2}(t-1)} + e^{-\frac{\delta}{2}t} = (2e^{\frac{\delta}{2}} + 1)e^{-\frac{\delta}{2}t} = e^{-\Omega(t)}. \tag{3.69}$$

This verifies **Condition 2** - to a much faster rate than is in fact required. It remains to verify **Condition 3**. For $\delta > 0$,

$$\begin{aligned}&\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(\rho)) \\ &= \mathbb{P}_f\left(\delta < \frac{1}{\hat{\alpha}_t}a^{-1}\left(\frac{\rho}{b(\hat{\beta}_t)}\right) - \ln\left(\frac{1}{\hat{\alpha}_t}a^{-1}\left(\frac{\rho}{b(\hat{\beta}_t)}\right)\right) - 1 \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right) \\ &= \mathbb{P}_f\left(\frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta)) \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right) \\ &+ \mathbb{P}_f\left(\frac{\rho}{b(\hat{\beta}_t)} < a(\hat{\alpha}_t L^+(\delta)) \text{ and } \frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t)\right).\end{aligned} \tag{3.70}$$

The above bound can be simplified a great deal. In the second term, the conditions in fact contradict: since a is taken to be a decreasing function of α, and $L^+(\delta) > 1$ for $\delta > 0$, the probability is 0. In the first term, since $0 < L^-(\delta) < 1$ for $\delta > 0$, and a is decreasing, the conditions may be combined to yield

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(\rho)) = \mathbb{P}_f\left(\frac{\rho}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta))\right). \tag{3.71}$$

Let $\rho = s(f) - \varepsilon = a(\alpha)b(\beta) - \varepsilon$. It is convenient to take $\varepsilon = a(\alpha)b(\beta)\tilde{\varepsilon}$ with $0 < \tilde{\varepsilon} < 1$, so $\rho = a(\alpha)b(\beta)(1-\tilde{\varepsilon})$. Recall that b is non-decreasing, and $\beta \le \hat{\beta}_t$. Hence,

$$\begin{aligned}\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f) - \varepsilon)) &= \mathbb{P}_f\left(\frac{a(\alpha)b(\beta)(1-\tilde{\varepsilon})}{b(\hat{\beta}_t)} > a(\hat{\alpha}_t L^-(\delta))\right) \\ &\le \mathbb{P}_f\left(a(\alpha)(1-\tilde{\varepsilon}) > a(\hat{\alpha}_t L^-(\delta))\right) \\ &= \mathbb{P}_f\left(\frac{\alpha}{\hat{\alpha}_t} < \frac{\alpha}{a^{-1}\left(a(\alpha)(1-\tilde{\varepsilon})\right)}L^-(\delta)\right)\end{aligned} \tag{3.72}$$

Let $\sigma = \alpha/a^{-1}(a(\alpha)(1-\tilde{\varepsilon}))$, and note that by Condition on a, $0 < \sigma < 1$. Letting $G_t \sim$ Gamma$(t,1)$, we may apply Lemma 5 for

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f)-\varepsilon)) \leq \mathbb{P}_f\left(G_{t-1} < (t-1)\sigma L^-(\delta)\right) \leq \left(\sigma L^-(\delta) e^{1-\sigma L^-(\delta)}\right)^{t-1} \tag{3.73}$$

Noting that $L^-(\delta) - \ln L^-(\delta) - 1 = \delta$, we have $L^-(\delta)e = e^{L^-(\delta)-\delta}$, and

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f)-\varepsilon)) \leq \left(\sigma e^{L^-(\delta)(1-\sigma)-\delta}\right)^{t-1} \leq \left(\sigma e^{1-\sigma}\right)^{t-1} e^{-\delta(t-1)}. \tag{3.74}$$

The last step follows as $0 < L^-(\delta) < 1$ for $\delta > 0$. This verifies **Condition 3**, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$, and Eq. (3.20) follows from Eq. (3.22), the definition of $\mathbb{K}_f(\rho)$ for this model.

Proof. [of Lemma 5.] Let $Y_1, \ldots, Y_t$ be i.i.d. Exp(1) random variables, and let $G = Y_1 + \ldots + Y_t$. For $0 < \gamma^- < 1 < \gamma^+ < \infty$,

$$\begin{aligned}
&\mathbb{P}\left(G < \gamma^- t\right) \\
&= \mathbb{P}\left(e^{-\left(\frac{1}{\gamma^-}-1\right)G} > e^{-\left(\frac{1}{\gamma^-}-1\right)\gamma^- t}\right) \\
&= \mathbb{P}\left(e^{-(\frac{1}{\gamma^-}-1)G} > e^{-(1-\gamma^-)t}\right) \\
&\leq \frac{\mathbb{E}\left[e^{-(\frac{1}{\gamma^-}-1)G}\right]}{e^{-(1-\gamma^-)t}} = \frac{\prod_{s=1}^t \mathbb{E}\left[e^{-(\frac{1}{\gamma^-}-1)Y_s}\right]}{e^{-(1-\gamma^-)t}} = \frac{(\gamma^-)^t}{e^{-(1-\gamma^-)t}} = \left(\gamma^- e^{1-\gamma^-}\right)^t.
\end{aligned} \tag{3.75}$$

The result for $\mathbb{P}(G > \gamma^+ t)$ follows similarly.

Proof. [of Theorem 7.] It remains to verify **Conditions 2 & 3**. **Condition 2** now takes the following form:

$$\mathbb{P}_{f_S}\left(||\hat{S}_t| - |S|| > \delta\right) = o(1/t). \tag{3.76}$$

Observe the decomposition,

$$\mathbb{P}_{f_S}\left(||\hat{S}_t| - |S|| > \delta\right) = \mathbb{P}_{f_S}\left(|\hat{S}_t| > |S| + \delta\right) + \mathbb{P}_{f_S}\left(|\hat{S}_t| < |S| - \delta\right). \tag{3.77}$$

We have the following bound, almost surely, on the size of $\hat{S}_t$: Letting $\#S$ denote the number of disjoint intervals in S, $|\hat{S}_t| \leq |S| + 2\varepsilon_t \#S$. As this is almost sure, and $\varepsilon_t \to 0$ with t, the first term in the decomposition above is 0 for all sufficiently large t. To bound the other term, note that without loss of generality, we may take $\delta < |S|$. For notational convenience, let $\alpha = 1 - \delta/|S|$, and note that $0 < \alpha < 1$.

In the event that $|\hat{S}_t| < \alpha|S|$, there exists a set of ε_t-intervals of those that intersect S that both cover a total measure of $\alpha|S|$, and contain all t samples from f_S. The number of ε_t-intervals intersecting S is at most $\lceil |S|/\varepsilon_t \rceil + 2\#S$. The number of ε_t-intervals needed to cover an area of $\alpha|S|$ is $\lceil \alpha|S|/\varepsilon_t \rceil$. Noting that the f_S samples are independent, and each falls in a given set of $\alpha|S|$-covering ε_t-intervals with probability at most α, we have

$$\begin{aligned}
\mathbb{P}_{f_S}\left(|\hat{S}_t| < \alpha|S|\right) &\leq \binom{\lceil |S|d_t \rceil + 2\#S}{\lceil \alpha|S|d_t \rceil} \alpha^t \\
&\leq \left(e\frac{\lceil |S|d_t \rceil + 2\#S}{\lceil \alpha|S|d_t \rceil}\right)^{\lceil \alpha|S|d_t \rceil} \alpha^t \\
&\leq \left(e\frac{|S|d_t + 2\#S + 1}{\alpha|S|d_t}\right)^{d_t} \alpha^t \\
&= \left(1 + \frac{2\#S+1}{|S|d_t}\right)^{d_t} e^{d_t}\alpha^{t-d_t} = e^{O(d_t)}\alpha^{t-d_t}.
\end{aligned} \tag{3.78}$$

It follows from this and the previous analysis that $\mathbb{P}_{f_S}\left(||\hat{S}_t| - |S|| > \delta\right) = e^{-\Omega(t)}$ in fact, verifying **Condition 2**. To verify **Condition 3**, note

$$\mathbb{P}_{f_S}(\delta < \mathbb{K}_{\hat{f}_t}(s(f_S) - \varepsilon)) \leq \mathbb{P}_{f_S}\left(\delta < \ln\left((|S| - \varepsilon)/|\hat{S}_t|\right)\right) = \mathbb{P}_{f_S}\left(|\hat{S}_t| < (|S| - \varepsilon)\, e^{-\delta}\right). \tag{3.79}$$

The additional case in $\mathbb{K}_f$ may be dispensed with observing that $\delta > 0$. Taking $\varepsilon < |S|$, it is convenient to define $\tilde{\varepsilon} = 1 - \varepsilon/|S|$. In which case,

$$\mathbb{P}_{f_S}(\delta < \mathbb{K}_{\hat{f}_t}(s(f_S) - \varepsilon)) \leq \mathbb{P}_{f_S}\left(|\hat{S}_t| < |S|\,(1 - \tilde{\varepsilon})\, e^{-\delta}\right). \tag{3.80}$$

Applying the previously established bound therefore yields,

$$\mathbb{P}_{f_S}(\delta < \mathbb{K}_{\hat{f}_t}(s(f_S) - \varepsilon)) \leq e^{O(d_t)}(1 - \tilde{\varepsilon})^{t-d_t} e^{-\delta(t-d_t)}, \tag{3.81}$$

verifying **Condition 3**. Equation (3.25) follows from Eq. (3.27), the definition of $\mathbb{K}_f(\rho)$ for this model.

Proof. [of Theorem 8.] It remains to verify **Conditions 1-3**, and the continuity of s under **I**.

Note that if $\mathbf{I}(f_{a,b}, f_{\tilde{a},\tilde{b}}) < \delta$, it follows that

$$\begin{aligned} \tilde{a} &\leq a \\ b &\leq \tilde{b} \\ \tilde{b}-\tilde{a} &< (b-a)e^{\delta}. \end{aligned} \tag{3.82}$$

It follows that $0 \leq \tilde{b} - b < (b-a)(e^{\delta}-1)$ and $0 \leq a - \tilde{a} < (b-a)(e^{\delta}-1)$. From this, we may conclude that any function of $f \in \mathscr{F}$ that is continuous with respect to the parameters is continuous with respect to f under $\mathbf{I}$. It follows, given the assumptions on s, that s is continuous under $\mathbf{I}$. Again, we take in this case that $\nu = \mathbf{I}$. Note that the continuity of s with respect to b makes $\mathbb{K}_f(\rho)$ continuous with respect to ρ. This, and the previous analysis, verifies **Condition 1**.

To verify **Condition 2**, note that $a \leq \hat{a}_t \leq \hat{b}_t \leq b$. Hence, we have the following:

$$\mathbb{P}_f\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) = \mathbb{P}_f\left((b-a) > (\hat{b}_t - \hat{a}_t)e^{\delta}\right) = \mathbb{P}_f\left(e^{-\delta} > \frac{\hat{b}_t - \hat{a}_t}{b-a}\right) \tag{3.83}$$

Here, we utilize the following Lemma, characterizing the distribution of $\hat{a}_t, \hat{b}_t$:

Lemma 6 *For $t \geqslant 2, 0 < \lambda < 1$:*

$$\mathbb{P}_{f_{a,b}}\left(\frac{\hat{b}_t - \hat{a}_t}{b-a} < \lambda\right) = \left(t(1-\lambda)+\lambda\right)\lambda^{t-1} \leq (t+1)\lambda^{t-1}. \tag{3.84}$$

The proof is given following this one. Hence we see that

$$\mathbb{P}_f\left(\mathbf{I}(\hat{f}_t, f) > \delta\right) \leq (t+1)e^{-\delta(t-1)} = e^{-\Omega(t)}, \tag{3.85}$$

verifying **Condition 2**.

For **Condition 3**, note that

$$\begin{aligned} \mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(\rho)) &= \mathbb{P}_f\left(\delta < \min_{\hat{b}_t \leq \tilde{b}}\left\{\ln\left(\frac{\tilde{b}-\hat{a}_t}{\hat{b}_t - \hat{a}_t}\right) : s(\hat{a}_t, \tilde{b}) > \rho\right\}\right) \\ &= \mathbb{P}_f\left(\max_{\hat{b}_t \leq \tilde{b}}\left\{s(\hat{a}_t, \tilde{b}) : \ln\left(\frac{\tilde{b}-\hat{a}_t}{\hat{b}_t - \hat{a}_t}\right) \leq \delta\right\} < \rho\right) \\ &= \mathbb{P}_f\left(\max_{\hat{b}_t \leq \tilde{b}}\left\{s(\hat{a}_t, \tilde{b}) : \tilde{b} \leq \hat{a}_t + e^{\delta}(\hat{b}_t - \hat{a}_t)\right\} < \rho\right) \\ &= \mathbb{P}_f\left(s(\hat{a}_t, \hat{a}_t + e^{\delta}(\hat{b}_t - \hat{a}_t)) < \rho\right) \\ &\leq \mathbb{P}_f\left(s(a, a + e^{\delta}(\hat{b}_t - \hat{a}_t)) < \rho\right). \end{aligned} \tag{3.86}$$

Hence we have that

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f)-\varepsilon)) \leq \mathbb{P}_f\left(s(a, a+e^{\delta}(\hat{b}_t-\hat{a}_t)) < s(a,b)-\varepsilon\right). \tag{3.87}$$

Given the continuity of s, let $\tilde{\varepsilon} > 0$ be such that $s(a, b-\tilde{\varepsilon}) \geqslant s(a,b)-\varepsilon$.

$$\begin{aligned}\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f)-\varepsilon)) &\leq \mathbb{P}_f\left(s(a, a+e^{\delta}(\hat{b}_t-\hat{a}_t)) < s(a,b-\tilde{\varepsilon})\right) \\ &= \mathbb{P}_f\left(a+e^{\delta}(\hat{b}_t-\hat{a}_t) < b-\tilde{\varepsilon}\right) \\ &= \mathbb{P}_f\left(\frac{\hat{b}_t-\hat{a}_t}{b-a} < e^{-\delta}\left(1-\frac{\tilde{\varepsilon}}{b-a}\right)\right) \\ &\leq (t+1)e^{-\delta(t-1)}\left(1-\frac{\tilde{\varepsilon}}{b-a}\right)^{t-1} = e^{-\Omega(t)}e^{-\delta(t-1)}.\end{aligned} \tag{3.88}$$

This verifies **Condition 3**, with $d_t = 1$, producing a bound of the correct order. This in turn verifies the policy as optimal, taking $\tilde{d}(t) = 2$, and Eq. 3.31 follows from the definition of $\mathbb{K}_f(\rho)$ for this model.

Proof. [of Lemma 6.] Let $X_1, \ldots, X_t$ be i.i.d. Uniform$[0,1]$ random variables. Note that we may then take $\hat{a}_t = a+(b-a)\min_n X_n$, $\hat{b}_t = a+(b-a)\max_n X_n$. Hence,

$$\mathbb{P}_{f_{a,b}}\left(\frac{\hat{b}_t-\hat{a}_t}{b-a} < \lambda\right) = \mathbb{P}\left(\max_n X_n - \min_n X_n < \lambda\right) \tag{3.89}$$

Let $M = \max_n X_n$ and $m = \min_n X_n$. Note that, conditioned on m, $M-m$ is distributed like the maximum of $t-1$ many Uniform$[0,1-m]$ random variables. Let $Y_1, \ldots, Y_{t-1}$ be i.i.d. Uniform$[0,1]$ random variables, so we may take $M-m = (1-m)\max_s Y_s$.

$$\begin{aligned}\mathbb{P}(M-m<\lambda|m) &= \mathbb{P}\left((1-m)\max_s Y_s < \lambda|m\right) \\ &= \mathbb{1}\{1-m \leq \lambda\} + \frac{\lambda^{t-1}}{(1-m)^{t-1}}\mathbb{1}\{1-m > \lambda\}\end{aligned} \tag{3.90}$$

Note that m is distributed with a density of $t(1-x)^{t-1}$ for $x \in [0,1]$. From the above then

$$\begin{aligned}\mathbb{P}_{f_{a,b}}\left(\frac{\hat{b}_t-\hat{a}_t}{b-a} < \lambda\right) &= \mathbb{P}(M-m<\lambda) \\ &= \mathbb{E}\left[\mathbb{P}(M-m<\lambda|m)\right] \\ &= \mathbb{P}(1-\lambda \leq m) + \mathbb{E}\left[\frac{\lambda^{t-1}}{(1-m)^{t-1}}\mathbb{1}\{1-\lambda > m\}\right] \\ &= \lambda^t + t(1-\lambda)\lambda^{t-1}.\end{aligned} \tag{3.91}$$

The result follows immediately.

Lemma 7 *Let $U_t \sim \chi_t^2$, and Z be a standard normal. For $z \geqslant 0$, and $0 < u^- < 1 < u^+ < \infty$, the following bounds hold:*

$$\begin{aligned}
\mathbb{P}\left(U_t > u^+ t\right) &\leq \left(u^+ e^{1-u^+}\right)^{\frac{t}{2}} \\
\mathbb{P}\left(U_t < u^- t\right) &\leq \left(u^- e^{1-u^-}\right)^{\frac{t}{2}} \\
\mathbb{P}(Z > z) &\leq e^{-z^2/2}.
\end{aligned} \tag{3.92}$$

Proof. [of Lemma 7.] For the normal bound, note that for any $z > 0$,

$$\mathbb{P}(Z > z) = \mathbb{P}\left(e^{zZ} > e^{z^2}\right) \leq \mathbb{E}\left[e^{zZ}\right] e^{-z^2} = e^{z^2/2 - z^2} = e^{-z^2/2}. \tag{3.93}$$

For the χ_t^2 bounds, let $0 < u^- < 1 < u^+$, and let $Z_1, \ldots, Z_t$ be i.i.d. standard normal random variables. Let $U_t = \sum_{i=1}^t Z_i^2$. Observe that

$$\begin{aligned}
\mathbb{P}\left(U_t > u^+ t\right) &= \mathbb{P}\left(e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t} > e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)u^+ t}\right) \\
&= \mathbb{P}\left(e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t} > e^{(u^+ - 1)t/2}\right) \\
&\leq \mathbb{E}\left[e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)U_t}\right] e^{-(u^+ - 1)t/2} \\
&= \mathbb{E}\left[e^{\left(\frac{1}{2} - \frac{1}{2u^+}\right)Z^2}\right]^t e^{-(u^+ - 1)t/2} \\
&= \left(\sqrt{u^+}\right)^t e^{-(u^+ - 1)t/2}.
\end{aligned} \tag{3.94}$$

The result follows immediately as a rearrangement of the above. The result for $\mathbb{P}(U_t < u^- t)$ follows similarly.

Proof. [of Theorem 10.]

Let $\underline{f} \in \mathscr{F}_P^N$ be fixed, with f_i as a normal density with mean μ_i, variance σ^2. Let $\mu^* = \max_i \mu_i$. It is convenient to denote $\underline{T}_\pi(n) = \left(T_\pi^1(n), \ldots, T_\pi^N(n)\right)$. Let i be a sub-optimal arm, and note the following relationship:

$$\begin{aligned}
T_\pi^i(n+1) &= 2 + \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1) = i\} \\
&\leq 2 + n_1^i(n,\varepsilon) + n_2^i(n,\varepsilon) + n_3^i(n,\varepsilon) + n_4^i(n,\varepsilon),
\end{aligned} \tag{3.95}$$

where

$$n_1^i(n,\varepsilon) = \sum_{t=2N}^{n} \mathbb{1}\left\{\pi(t+1)=i; v_i(t,\underline{T}_\pi(t)) > \mu^* - \varepsilon, \hat{\mu}^i_{T^i_\pi(t)} \leq \mu_i + \varepsilon, \frac{\hat{\sigma}^2(\underline{T}_\pi(t))}{\sigma^2} \leq 1+\varepsilon\right\}, \tag{3.96}$$

and

$$\begin{aligned} n_2^i(n,\varepsilon) &= \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i; v_i(t,\underline{T}_\pi(t)) > \mu^* - \varepsilon, \hat{\mu}^i_{T^i_\pi(t)} > \mu_i + \varepsilon\}, \\ n_3^i(n,\varepsilon) &= \sum_{t=2N}^{n} \mathbb{1}\left\{\pi(t+1)=i; v_i(t,\underline{T}_\pi(t)) > \mu^* - \varepsilon, \frac{\hat{\sigma}^2(\underline{T}_\pi(t))}{\sigma^2} > 1+\varepsilon\right\}, \\ n_4^i(n,\varepsilon) &= \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i; v_i(t,\underline{T}_\pi(t)) \leq \mu^* - \varepsilon\}. \end{aligned} \tag{3.97}$$

The proof proceeds by bounding the expectation of each of the four terms. Let $\varepsilon > 0$ be such that $2\varepsilon < \Delta_j = \mu^* - \mu_j$ for each sub-optimal j.

To bound $n_1^i(n,\varepsilon)$, note

$$\begin{aligned} n_1^i(n,\varepsilon) &\leq \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i; \mu_i + \varepsilon + \frac{\sigma\sqrt{1+\varepsilon}}{\sqrt{T^i_\pi(t)-1}}\sqrt{(t-N)\left(t^{\frac{2}{t-N}}-1\right)} > \mu^* - \varepsilon\} \\ &= \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i; \frac{\sigma^2(1+\varepsilon)}{(\Delta_i - 2\varepsilon)^2}(t-N)\left(t^{\frac{2}{t-N}}-1\right)+1 > T^i_\pi(t)\} \\ &\leq \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i; \frac{\sigma^2(1+\varepsilon)}{(\Delta_i - 2\varepsilon)^2}(n-N)\left(n^{\frac{2}{n-N}}-1\right)+1 > T^i_\pi(t)\}. \end{aligned} \tag{3.98}$$

The last step follows, as the function inside the indicator is an increasing function of t over the indicated ranges. It follows then that

$$n_1^i(n,\varepsilon) \leq 4 + \frac{\sigma^2(1+\varepsilon)}{(\Delta_i - 2\varepsilon)^2}(n-N)\left(n^{\frac{2}{n-N}}-1\right). \tag{3.99}$$

This follows from the previous bound, viewing $T^i_\pi(n)$ as a sum of indicators $\mathbb{1}\{\pi(n)=i\}$, and seeing that the conditioning restricts how many of these indicators in the above sum can be non-zero - the $+4$ catches the $\pi(n+1)=i$ indicator term that is not included in the sum for $T^i_\pi(n)$, and the two pulls of arm i that occurred for $t \leq 2N$. Note, this bound is almost sure.

To bound $n_2^i(n,\varepsilon)$, note

$$\begin{aligned}
n_2^i(n,\varepsilon) &\leq \sum_{t=2N}^{n} \mathbb{1}\{\pi(t+1)=i;\hat{\mu}^i_{T^i_\pi(t)} > \mu_i+\varepsilon\} \\
&= \sum_{t=2N}^{n}\sum_{k_i=2}^{t} \mathbb{1}\{\pi(t+1)=i;\hat{\mu}^i_{k_i} > \mu_i+\varepsilon, T^i_\pi(t)=k_i\} \\
&= \sum_{t=2N}^{n}\sum_{k_i=2}^{t} \mathbb{1}\{\pi(t+1)=i;T^i_\pi(t)=k_i\}\mathbb{1}\{\hat{\mu}^i_{k_i} > \mu_i+\varepsilon\} \qquad (3.100) \\
&\leq \sum_{k_i=2}^{n} \mathbb{1}\{\hat{\mu}^i_{k_i} > \mu_i+\varepsilon\}\sum_{t=k_i}^{n} \mathbb{1}\{\pi(t+1)=i;T^i_\pi(t)=k_i\} \\
&\leq \sum_{k_i=2}^{n} \mathbb{1}\{\hat{\mu}^i_{k_i} > \mu_i+\varepsilon\}.
\end{aligned}$$

The last inequality follows as, for fixed k_i, $\pi(t+1)=i;T^i_\pi(t)=k_i$ may be true for at most one value of t, sample-path-wise. Noting that $\hat{\mu}^i_k \sim \mu_i + (\sigma/\sqrt{k})Z$ for Z a standard normal random variable, we have that

$$\begin{aligned}
\mathbb{E}_{\underline{f}}\left[n_2^i(n,\varepsilon)\right] &\leq \sum_{k_i=2}^{n} \mathbb{P}_{\underline{f}}\left(\hat{\mu}^i_{k_i} > \mu_i+\varepsilon\right) \\
&= \sum_{k_i=2}^{n} \mathbb{P}\left(Z > \frac{\varepsilon}{\sigma}\sqrt{k_i}\right) \\
&\leq \sum_{k_i=1}^{\infty} e^{-k_i(\varepsilon/\sigma)^2/2} \qquad (3.101) \\
&= \frac{1}{e^{(\varepsilon/\sigma)^2/2}-1}.
\end{aligned}$$

This follows from the bound of Lemma 7 on the tail probabilities for standard normals.

To bound $n_3^i(n,\varepsilon)$, note

$$n_3^i(n,\varepsilon) \leq \sum_{t=2N}^{n} \mathbb{1}\{\hat{\sigma}^2(\underline{T}_\pi(t)) > \sigma^2(1+\varepsilon)\}, \qquad (3.102)$$

hence

$$\begin{aligned}
\mathbb{E}_{\underline{f}}\left[n_3^i(n,\varepsilon)\right] &\leq \sum_{t=2N}^{n} \mathbb{P}_{\underline{f}}\left(\hat{\sigma}^2(\underline{T}_\pi(t)) > \sigma^2(1+\varepsilon)\right) \\
&= \sum_{t=2N}^{n} \mathbb{E}_{\underline{f}}\left[\mathbb{P}_{\underline{f}}\left(\hat{\sigma}^2(\underline{T}_\pi(t)) > \sigma^2(1+\varepsilon)\middle| T^1_\pi(t),\ldots,T^N_\pi(t)\right)\right]. \qquad (3.103)
\end{aligned}$$

At this point, note that for each i, $\hat{\sigma}^2_i(k_i) \sim \sigma^2 U^i_{k_i-1}/k_i$ where U^i_d is a χ^2-random variable of degree d. And as the arms are independent, we have

$$\hat{\sigma}^2(\underline{k}) \sim \frac{\sum_{i=1}^N \sigma^2 U^i_{k_i-1}}{\sum_{i=1}^N (k_i-1)} \sim \frac{\sigma^2}{\sum_{i=1}^N k_i - N} U_{\sum_{i=1}^N k_i - N}, \qquad (3.104)$$

again taking U_d as χ^2 with degree d. Noting that $\sum_{i=1}^{N} T_\pi^i(t) = t$, always, we have that

$$
\begin{aligned}
\mathbb{E}_{\underline{f}}\left[n_3^i(n,\varepsilon)\right] &\leq \sum_{t=2N}^{n} \mathbb{E}_{\underline{f}}\left[\mathbb{P}_{\underline{f}}\left(\frac{\sigma^2}{t-N}U_{t-N} > \sigma^2(1+\varepsilon)\middle|T_\pi^1(t),\ldots,T_\pi^N(t)\right)\right] \\
&= \sum_{t=2N}^{n} \mathbb{P}\left(\frac{\sigma^2}{t-N}U_{t-N} > \sigma^2(1+\varepsilon)\right) \\
&= \sum_{t=2N}^{n} \mathbb{P}\left(U_{t-N} > (1+\varepsilon)(t-N)\right) \qquad (3.105) \\
&\leq \sum_{t=1}^{\infty} \mathbb{P}\left(U_t > (1+\varepsilon)t\right) \\
&\leq \sum_{t=1}^{\infty} \left((1+\varepsilon)e^{-\varepsilon}\right)^{t/2}.
\end{aligned}
$$

The last step utilizes the Chernoff bound of Lemma 7 on χ^2-tail probabilities. It follows from this that:

$$
\mathbb{E}_{\underline{f}}\left[n_3^i(n,\varepsilon)\right] \leq \frac{\sqrt{(1+\varepsilon)e^{-\varepsilon}}}{1-\sqrt{(1+\varepsilon)e^{-\varepsilon}}}. \tag{3.106}
$$

To bound $n_4^i(n,\varepsilon)$, note that in the event that $\pi(t+1) = i$, from the structure of the policy it must be true that $v_i(t,\underline{T}_\pi(t)) = \max_j v_j(t,\underline{T}_\pi(t))$. Hence, in the event that $v_i(t,\underline{T}_\pi(t)) \leq \mu^* - \varepsilon$, it must also be true that for any optimal arm i^*, $v_{i^*}(t,\underline{T}_\pi(t)) \leq \mu^* - \varepsilon$. Hence the following bounds:

$$
\begin{aligned}
n_4^i(n,\varepsilon) &\leq \sum_{t=2N}^{n} \mathbb{1}\left\{\pi(t+1) = i; v_{i^*}(t,\underline{T}_\pi(t)) \leq \mu^* - \varepsilon\right\} \\
&\leq \sum_{t=2N}^{n} \mathbb{1}\left\{\hat{\mu}^{i^*}_{T_\pi^{i^*}(t)} + \frac{\hat{\sigma}(\underline{T}_\pi(t))}{\sqrt{T_\pi^{i^*}(t)-1}}\sqrt{(t-N)\left(t^{\frac{2}{t-N}}-1\right)} \leq \mu^* - \varepsilon\right\},
\end{aligned} \tag{3.107}
$$

hence,

$$
\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] \leq \sum_{t=2N}^{n} \mathbb{P}_{\underline{f}}\left(\hat{\mu}^{i^*}_{T_\pi^{i^*}(t)} + \frac{\hat{\sigma}(\underline{T}_\pi(t))}{\sqrt{T_\pi^{i^*}(t)-1}}\sqrt{(t-N)\left(t^{\frac{2}{t-N}}-1\right)} \leq \mu^* - \varepsilon\right). \tag{3.108}
$$

Performing the same conditioning on $(T_\pi^1(t),\ldots,T_\pi^N(t))$ as in Eq. (3.103), note that $\hat{\mu}^{i^*}_{T_\pi^{i^*}(t)}$ and $\hat{\sigma}(\underline{T}_\pi(t))$ are in fact independent of each other, given the underlying normal distributions of the arm pulls - $\hat{\mu}^{i^*}_{k_i}$ is independent of $\hat{\sigma}_i^2(k_i)$, and sample variances of the other arms combined in $\hat{\sigma}^2(\underline{k})$ are independent of each other. Further, as noted previously, $\hat{\sigma}^2(\underline{T}_\pi(t)) \sim \sigma^2/(t-N)U_{t-N}$, and $\hat{\mu}^{i^*}_{T_\pi^{i^*}(t)} \sim \mu^* - (\sigma/\sqrt{T_\pi^{i^*}(t)})Z$ (the $-$ sign is convenient for the manupations to follow), with

U_{t-N}, Z as independent χ^2_{t-N} and standard normal random variables, respectively. Hence,

$$\begin{aligned}
&\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] \\
&\leq \sum_{t=2N}^{n} \mathbb{P}_{\underline{f}}\left(-\frac{\sigma}{\sqrt{T_\pi^{i^*}(t)}}Z + \frac{\sigma}{\sqrt{t-N}}\frac{\sqrt{U_{t-N}}}{\sqrt{T_\pi^{i^*}(t)-1}}\sqrt{(t-N)\left(t^{\frac{2}{t-N}}-1\right)} \leq -\varepsilon\right) \\
&= \sum_{t=2N}^{n} \mathbb{P}_{\underline{f}}\left(\frac{\varepsilon}{\sigma} + \frac{\sqrt{U_{t-N}}}{\sqrt{T_\pi^{i^*}(t)-1}}\sqrt{t^{\frac{2}{t-N}}-1} \leq \frac{Z}{\sqrt{T_\pi^{i^*}(t)}}\right).
\end{aligned} \tag{3.109}$$

Given that it must be that, for any t, $T_\pi^{i^*}(t) \in \{3,4,\ldots,t\}$, we may apply a union bound to the above, yielding

$$\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] \leq \sum_{t=2N}^{n}\sum_{s=2}^{t} \mathbb{P}\left(\frac{\varepsilon}{\sigma}\sqrt{s} + \frac{\sqrt{s}}{\sqrt{s-1}}\sqrt{U_{t-N}}\sqrt{t^{\frac{2}{t-N}}-1} \leq Z\right). \tag{3.110}$$

Note, the $\underline{f}$ subscript has been dropped, as the remaining random variables no longer depend in distribution on the $\{\mu_j\}$ or σ. It remains to bound the above double sum.

Note the following bounds, for $\alpha, \beta \geqslant 0$:

$$\begin{aligned}
\mathbb{P}\left(\alpha\sqrt{s} + \beta\sqrt{U_k/k} \leq Z\right) &\leq \mathbb{E}\left[\exp\left(-\frac{1}{2}\left(\alpha^2 s + 2\alpha\beta\sqrt{s}\sqrt{U_k/k} + \beta^2(U_k/k)\right)\right)\right] \\
&\leq e^{-\alpha^2 s/2}\mathbb{E}\left[\exp\left(-\frac{1}{2}\left(\beta^2(U_k/k)\right)\right)\right] \\
&= e^{-\alpha^2 s/2}\left(1+\frac{\beta^2}{k}\right)^{-k/2}.
\end{aligned} \tag{3.111}$$

Applying this to the above,

$$\begin{aligned}
\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] &\leq \sum_{t=2N}^{n}\sum_{s=2}^{t} \frac{e^{-(\varepsilon/\sigma)^2 s/2}}{\left(1+\left(\frac{s}{s-1}\right)\left(t^{\frac{2}{t-N}}-1\right)\right)^{\frac{t-N}{2}}} \\
&= \sum_{t=2N}^{n}\frac{1}{t}\sum_{s=2}^{t} \frac{e^{-(\varepsilon/\sigma)^2 s/2}}{\left(1+\frac{1}{s-1}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} \\
&\leq \sum_{t=2N}^{n}\frac{1}{t}\sum_{s=2}^{\infty} \frac{e^{-(\varepsilon/\sigma)^2 s/2}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}}.
\end{aligned} \tag{3.112}$$

It is proven following this proof, as Lemma 8, that for any t, the s-sum is $O(1/\ln t)$, hence we have that

$$\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] \leq \sum_{t=2N}^{n}\frac{1}{t}O\left(\frac{1}{\ln t}\right) = O(\ln\ln n). \tag{3.113}$$

Combining the four terms, we have from Eq. (3.95) (noting that $T_\pi^i(n) \leq T_\pi^i(n+1)$) that

$$\begin{aligned}\mathbb{E}_{\underline{f}}\left[T_\pi^i(n)\right] &\leq 2+\mathbb{E}_{\underline{f}}\left[n_1^i(n,\varepsilon)\right]+\mathbb{E}_{\underline{f}}\left[n_2^i(n,\varepsilon)\right]+\mathbb{E}_{\underline{f}}\left[n_3^i(n,\varepsilon)\right]+\mathbb{E}_{\underline{f}}\left[n_4^i(n,\varepsilon)\right] \\ &= \frac{\sigma^2(1+\varepsilon)}{\left(\Delta_i-2\varepsilon\right)^2}(n-N)\left(n^{\frac{2}{n-N}}-1\right)+O(\ln\ln n)+O(1).\end{aligned} \tag{3.114}$$

It follows that for each sub-optimal i,

$$\limsup_n \frac{\mathbb{E}_{\underline{f}}\left[T_\pi^i(n)\right]}{\ln n} \leq \frac{\sigma^2(1+\varepsilon)}{\left(\Delta_i-2\varepsilon\right)^2}\limsup_n \frac{(n-N)\left(n^{\frac{2}{n-N}}-1\right)}{\ln n} = \frac{2\sigma^2(1+\varepsilon)}{\left(\Delta_i-2\varepsilon\right)^2}. \tag{3.115}$$

Summing this bound over all sub-optimal i, we have the following bounds on total expected mistakes,

$$\limsup_n \frac{M_\pi^{\underline{f}}(n)}{\ln n} \leq \sum_{i:\mu_i\neq\mu^*} \frac{2\sigma^2(1+\varepsilon)}{\left(\Delta_i-2\varepsilon\right)^2} \leq \sum_{i:\mu_i\neq\mu^*} \frac{2\sigma^2}{\Delta_i^2}. \tag{3.116}$$

To complete the proof, it is enough to observe the lower bound bound provided by Theorem 3, as in Eq. (2.27).

Lemma 8 *For $\alpha > 0$,*

$$\sum_{s=2}^{\infty} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} = O\left(\frac{1}{\ln t}\right). \tag{3.117}$$

Proof. [of Lemma 8.] Let $\nu(t)$ be an increasing function of t. For t such that $\nu(t) > 2$, we have

$$\begin{aligned}&\sum_{s=2}^{\infty} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} \\ &= \sum_{s=2}^{\lfloor \nu(t)\rfloor} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} + \sum_{s=\lceil \nu(t)\rceil}^{\infty} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} \\ &\leq \sum_{s=2}^{\lfloor \nu(t)\rfloor} \frac{1}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} + \sum_{s=\lceil \nu(t)\rceil}^{\infty} e^{-\alpha s} \\ &\leq \frac{\lfloor \nu(t)\rfloor}{\left(1+\frac{1}{\lfloor \nu(t)\rfloor}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} + \frac{e^{-\alpha\lceil \nu(t)\rceil}}{1-e^{-\alpha}} \\ &\leq \frac{\nu(t)}{\left(1+\frac{1}{\nu(t)}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} + \frac{e^{-\alpha \nu(t)}}{1-e^{-\alpha}}.\end{aligned} \tag{3.118}$$

For all t sufficiently large, we have $1-t^{-2/(t-N)} \geqslant (\ln t)/(t-N)$. Applying this to the above, and taking $\nu(t) = (\ln t)^p$ for $0 < p < 1$, we have

$$\begin{aligned}\sum_{s=2}^{\infty} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} &\leq \frac{(\ln t)^p}{\left(1+\frac{(\ln t)^{1-p}}{t-N}\right)^{\frac{t-N}{2}}} + \frac{e^{-\alpha(\ln t)^p}}{1-e^{-\alpha}} \\ &\leq O(1)\left((\ln t)^p e^{-\frac{1}{2}(\ln t)^{1-p}} + e^{-\alpha(\ln t)^p}\right)\end{aligned} \tag{3.119}$$

Note, the bound on the first term can be derived from the asymptotics of

$$\frac{t-N}{2}\ln\left(1+\frac{(\ln t)^{1-p}}{t-N}\right). \tag{3.120}$$

As the exponential function grows faster than any polynomial of its argument, the above bounds can in fact be utilized to prove that for any $\delta > 0$,

$$\sum_{s=2}^{\infty} \frac{e^{-\alpha s}}{\left(1+\frac{1}{s}\left(1-t^{-\frac{2}{t-N}}\right)\right)^{\frac{t-N}{2}}} = o\left(\frac{1}{(\ln t)^{\delta}}\right). \tag{3.121}$$

Chapter 4

Halting Bandits

The models of the previous two chapters have a number of significant restrictions, which will be explored to some extent by a new model in this chapter. In particular, in the previous chapters: i) of the arms the controller faced, one or more were identified as ‘best’ and remained so through all time - the goal was pulling these as frequently as possible, ii) the actual results of the arm pulling did not matter in themselves (or for the goal), they only served to inform future arm pulling decisions.

In this chapter, the controller wishes to maximize some utility function of the actual results returned by the arms pulled, for instance the total expected reward returned by all pulls over some time horizon - slot machines being the classic example. Because the utility function depends on the results of each arm pull, and because the results for a given arm may evolve as that arm is repeatedly pulled - the classic example being an arm slowly breaking down as it is used - the idea of a constant ‘best’ arm to pull is no longer applicable. At any time, the best arm to pull next will depend on the current state of each arm, and what is known to the controller about the future trajectories of each arm.

In particular, each arm is taken to return a sequence of (potentially stochastic) rewards or losses as it is pulled. We consider the problem of maximizing several utility functions of the total rewards gained or losses incurred by the controller’s arm-pulling, over some time horizon defined by the first arm to break down (potentially randomly) from overuse. Note, this adds an interesting element to the classic ‘exploration vs exploitation’ dilemma previously mentioned, as the choice to pull a given arm may (through breaking that arm) exclude possible future exploitation of other arms! The value of a given arm pull cannot therefore be considered purely in terms of the reward or loss associated with that pull, the effects of the decision itself must also be considered. For the utility functions considered, we derive a formula for considering the 'true value' of the decision to pull an arm, as a function of what is currently known about that arm, and show that the optimal policy for maximizing the expected utility is myopic, in the sense of always pulling the arm with the current largest 'value'.

4.1 Formulation and Prior Work

The formulation of the problem in this chapter differs in several key ways from that of the previous, and as such it is worth rehashing some established notation while additionally giving some new. In particular, the primary uncertainty associated with each arm in the previous chapters regarded the true underlying dynamics of the arm processes the controller faced. In this chapter, the underlying laws are taken to be known in advance - the primary uncertainty therefore is simply due to the stochastic evolution of the arm processes themselves.

A controller is presented with a finite collection of $N \geqslant 2$ filtered probability spaces, $(\Omega^i, \mathscr{F}^i, \mathbb{P}^i, \mathbb{F}^i)$, for $1 \leq i \leq N < \infty$, representing N environments in which experiments will be performed or rewards collected - the 'bandits'. To each space, we associate an $\mathbb{F}^i$-adapted *reward process* $X^i = \{X^i_t\}_{t \geqslant 0}$. For $t \in \{0, 1, \ldots\}$, we take $X^i_t (= X^i_t(\omega^i)) \in \mathbb{R}$ to represent the reward available from the i^{th} bandit on its t^{th} activation. We restrict the reward process of each bandit in the following way, that

$$\mathbb{E}^i \left[\sup_{n \geqslant 0} |X^i_n| \right] < \infty. \tag{4.1}$$

We denote the collection of reward processes as $\mathbb{X}$.

Remark 5. In this chapter, unlike the previous, it is notationally convenient in terms of the formulas involved to take the arm processes as starting at time $t = 0$. Note, because the reward processes are taken to be $\mathbb{F}^i$-adapted, at the conclusion of the first t pulls (representing rewards $X^i_0, \ldots, X^i_{t-1}$), the value X^i_t is determined, and known to the controller. Future rewards, however, are only known to the extent allowed by conditioning on the available information.

Additionally, to each bandit, we associate an $\mathbb{F}^i$-stopping time $\sigma^i > 0$, the 'halting time' of the bandit. At time σ^i, we take the bandit to be stopped, and no longer capable of being activated. We take the following restriction on σ^i, that $\mathbb{P}^i(\sigma^i < \infty) = 1$ and that for all $t < \sigma^i$, we have almost surely that

$$\mathbb{P}^i(\sigma^i = t+1 | \mathscr{F}^i(t)) > 0. \tag{4.2}$$

That is, at every point in time t prior to stopping, there is a positive probability of halting in the next round. In the models we consider, the controller will activate bandits until the first time some bandit halts, at which point a reward based on the final state of each bandit will be collected. Hence we refer to these as ‘single payout’ bandits, as rewards are assigned at only one time, rather than collected cumulatively as in the previous chapters.

We embed these bandits in a larger ‘global’ probability space

$$(\Omega, \mathscr{G}, \mathbb{P}) = \left(\otimes_{i=1}^{N} \Omega^i, \otimes_{i=1}^{N} \mathscr{F}^i, \otimes_{i=1}^{N} \mathbb{P}^i\right),$$

a standard product-space construction, representing the environment of the controller - aware information from all bandits. This model captures the first key assumption: *the bandits are mutually independent* (e.g., X^i, X^j are independent relative to $\mathbb{P}$ for $i \neq j$). Expectations relative to the local space, i.e., bandit i, will be denoted $\mathbb{E}^i$, while expectations relative to the global space are simply $\mathbb{E}$.

In what follows, we reserve the term ‘round’ to differentiate global, controller time, denoted with s, from local bandit times, denoted by t. In each round, the controller activates a bandit, advancing its local time by one time step. All bandits begin at local time 0, and advance only on activation. This is the second key assumption, that in every round *unactivated bandits remain frozen*. This proceeds until one of the bandits halts, which in turn halts the control process. At this round, the controller receives a reward that is a function of the current state of each bandit - again, justifying the descriptor ‘single payout’.

The controller needs a *control policy* π, a stochastic process on $(\Omega, \mathscr{G}, \mathbb{P})$ that specifies, at each round s of global time, which bandit to activate and collect from, e.g., $\pi(s)(= \pi(s, \omega)) = i$ activates bandit i at round s. We restrict ourselves to the set of policies $\mathscr{P}$ defined to be *non-anticipatory*. A policy π is non-anticipatory if the choice of which bandit to activate at round s does not depend on outcomes that have not yet occurred, or information not yet available.

Remark 6. We adopt the following notational liberty, allowing a random variable Z defined on a local space Ω^i to also be considered as a random variable on the global space Ω, taking

$Z(\omega) = Z(\omega^i)$, where $\omega = (\omega^1, \ldots, \omega^N) \in \Omega$. Via this extension, we may take expectations involving a process X^i, or $\mathbb{F}^i$-stopping times, relative to $\mathbb{P}$ or $\mathbb{P}^i$, without additional notational overhead.

Given a policy π, it is convenient to be able to translate between global time and local time. Define $S^i_\pi(t)$ to be *the round at which bandit i is activated for the t^{th} time* when the controller operates according to policy π. This may be expressed as

$$\begin{aligned} S^i_\pi(0) &= \inf\{s \geqslant 0 : \pi(s) = i\}, \\ S^i_\pi(t+1) &= \inf\{s > S^i_\pi(t) : \pi(s) = i\}. \end{aligned} \tag{4.3}$$

Utilizing this notation, we may define a *global halting time σ_π*, i.e., the first round under policy π at which one of the bandits has halted, ending the control process:

$$\sigma_\pi = \min_i \{S^i_\pi(\sigma^i - 1)\} + 1. \tag{4.4}$$

We may also define $T^i_\pi(s)$ denote the *local time of bandit i just prior to the s^{th} round under a policy π*, i.e., $T^i_\pi(0) = 0$, and for $s > 0$,

$$T^i_\pi(s) = \sum_{s'=0}^{s-1} \mathbb{1}\{\pi(s') = i\}. \tag{4.5}$$

It is convenient to define the global time analog, $T_\pi(s) = T^{\pi(s)}_\pi(s)$ to denote the current local time of the bandit activated at round s under policy π. This will allow us to define concise global time analogs of several processes. For instance, we define the *global reward process X_π* on $(\Omega, \mathscr{G}, \mathbb{P})$ as

$$X_\pi(s) = X^{\pi(s)}_{T_\pi(s)},$$

giving the reward available from collection $\mathbb{X}$ under policy π at round s.

In what follows, for a given policy π, we take the final reward the controller receives to be a function of the final state of the bandits, generally a linear combination of $\{X^i_{T^i_\pi(\sigma_\pi)}\}_{1 \leq i \leq N}$. The most important model considered will be the problem of maximizing the total collective payout on halting,

$$\sum_{i=1}^{N} \mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} | \mathscr{G}_0\right]. \tag{4.6}$$

To maximize her expected reward, in every round the controller's decision of which bandit to activate must balance not only the current state of each bandit, but also the probability of halting that bandit and in doing so ending the process, and losing all potential future rewards.

The remainder of the chapter proceeds in the following way: as a starting point, we develop an optimal policy for maximizing expected total reward in the case that the bandits are known to have non-increasing rewards, and payout is received only from the bandit that halts; we then consider a model of general reward processes, under which payout is received at the time of first halting, based on the states of all bandits; we develop an optimal policy for maximizing expected payout under this 'collective payout' model by transforming it to an equivalent instance of the previous 'monotone, solo-payout' model; we then utilize this collective payout model to solve for optimal policies under additional payout models, based on singling out bandits that did or did not halt, i.e., costs are incurred for bandits that did not halt, while rewards are collected from those that did (or vice versa).

4.1.1 Global Information vs. Local Information

One of the intricacies of the results to follow is in properly distinguishing and determining what information is available to the controller to act on at a given time. For each bandit i, the filtration $\mathbb{F}^i = \{\mathscr{F}^i(t)\}_{t \geqslant 0}$ represents the progression of information available about that bandit - the σ-algebra $\mathscr{F}^i(t)$ representing the local information available about bandit i at local time t, such as the process history of X^i. Taking X^i as $\mathbb{F}^i$-adapted as we do, we have $\sigma(X^i_0, X^i_1, \dots, X^i_t) \subset \mathscr{F}^i(t)$.

At round s, the total, global information available to the controller is determined by the state of each bandit at that round, i.e. acting under a given policy π until round s, the global information available at round s is given by the σ-algebra $\bigotimes_{i=1}^N \mathscr{F}^i(T^i_\pi(s))$. We may therefore refine the prior definition of non-anticipatory policies to be the set of policies $\mathscr{P}$ such that for each $s \geqslant 0$, $\pi(s)$ is measurable with respect to the prior σ-algebra, i.e., determined by the information available at round s. Weaker definitions of non-anticipatory, such as allowing dependence on random events, e.g., coin flips, are addressed in Section 4.3.4. It is convenient to define the initial global σ-algebra $\mathscr{G}_0 = \bigotimes_{i=1}^N \mathscr{F}^i(0)$, representing the initial information available from each bandit, which is independent of policy π.

Additionally, given a policy π, it is necessary to define a set of policy-dependent filtrations in the following way: let $\mathbb{H}^i_\pi = \{\mathscr{H}^i_\pi(t)\}_{t \geqslant 0}$, where $\mathscr{H}^i_\pi(t) = \bigotimes_{j=1}^N \mathscr{F}^j(T^j_\pi(S^i_\pi(t)))$ represents the total information available to the controller about all bandits, prior to the t^{th} activation of bandit i under π. It is indexed by the local time of bandit i, but at each time t gives the current state of information of each bandit. Note that, since $T^i_\pi(S^i_\pi(t)) = t$, $\mathscr{H}^i_\pi(t)$ contains the information available in $\mathscr{F}^i(t)$. This filtration is necessary for expressing local stopping times, i.e., concerning X^i, from the perspective of the controller - $\mathbb{F}^i$-stopping times no longer suffice, since the controller has access to information from all the other processes as well. Note though, $\mathbb{F}^i$-stopping times may be viewed as $\mathbb{H}^i_\pi$-stopping times, cf. Remark 4.1. Ultimately, the optimal policy result demonstrates that any decision about a given bandit depends only on information from that bandit, thus rendering these filtrations unnecessary in practice. However, they are a technical necessity for the proof of that result.

When discussing stopping times, we will utilizing the following notation: For a general filtration $\mathbb{J}$ (e.g., $\mathbb{J} = \mathbb{F}^i, \mathbb{H}^i_\pi$), we denote by $\hat{\mathbb{J}}(t)$ the set of all $\mathbb{J}$-stopping times strictly greater than t ($\mathbb{P}^i, \mathbb{P}$-a.e.). For a $\mathbb{J}$-stopping time τ, $\hat{\mathbb{J}}(\tau)$ is similarly defined.

4.1.2 Prior Work

Consider a model in the above framework, in which rewards are collected on every activation, rather than at the time of halting. In that case, the value function may be written as

$$V_\pi(\mathbb{X}) = \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^{T^i_\pi(\sigma_\pi)-1} X^i_t \middle| \mathscr{G}_0\right], \tag{4.7}$$

taking empty sums to be equal to 0. Comparing the above with the value function described in Eq. (2.4), this model can be seen as a natural extension of the model considered in Section 2.1.1 (which inspired the models of the previous two chapters) to a bandit-defined time horizon through this halting mechanism. Framing the problem in terms of ‘single payouts’ instead of cumulative, as in the above value function, both simplifies the presentation of the results to follow, and allows for convenient extension to alternative payout models as well.

With this cumulative payout model in mind, however, the inspiration for the model considered in this chapter largely comes from Gittins and Jones [25] (see [26] for a modern treatment), who considered the problem of maximizing the expected present value of the total reward collected from finite state Markov chain bandits with a constant discount factor, i.e., maximizing total expected (discounted) reward over an infinite time horizon. This seminal work provided a dynamic allocation index policy for directing bandit activations, based on a formula that may be computed for each bandit depending only on the state of that bandit, effectively decoupling the bandits from each other in the decision process. The derivation of this optimal index policy in that work depended on an intricate exchange argument, considering 'policy improvement' by exchanging activations of various bandits under a given policy according to the so called 'Gittins index'. Many works followed this (see [23] for a general overview), providing alternative proofs of the same results - worth mention in particular is Whittle [77], which provided the interpretation of the Gittins index of a bandit in terms of the equivalent lump sum worth permanently abandoning a bandit for.

It will be shown in the sections to follow that the model presented here subsumes and extends the classical Gittins model of maximizing total expected discounted reward over an infinite horizon. Some proofs of the Gittins result also depend on various restrictions of the underlying bandit processes, for instance taking them to not only be Markov chains, but having finite state space as well [72]. The work presented here represents a considerable extension in this case, generalizing to essentially arbitrary reward processes, limited only in not admitting infinite expected rewards. Most directly, this present work is an extension of the results of Cowan and Katehakis [15] in the Gittins model, the results of which extended the Gittins model to sequences of non-uniform discounts and arbitrary depreciation, and [18] which considered a similar halting model in the case of Markov chains. Additionally, in these prior works, the bandits were generally treated as of equal importance in the reward collection process. In this work, we use the mechanism of halting to single out certain bandits - for instance, viewing halted bandits as 'successful' and un-halted bandits as unsuccessful - and in doing may treat payout models beyond the scope of the original Gittins formulation.

We find the derivation and proof presented here of the optimal index policy particularly satisfying, in terms of both clarity and intuition. In particular, as will become clear, the relative simplicity of the proof suggests the importance of viewing the decision process not as decisions over ‘space’, e.g., solving for the optimal decision for each potential state of the bandits, but rather as decisions over ‘time’, e.g., viewing the decision to activate a given bandit in terms of a duration of activation. Additionally, we view the correspondence established between the solo-payout (rewards collected only from the bandit that halts) and collective-payout (rewards collected from all bandits on halting) models as illustrative of why the optimal decision process in the Gittins formulation decomposes into treating the bandits individually, through the computation of the indices.

4.2 Maximizing Solo Payouts: Non-Increasing Rewards

In this section, we consider the problem of maximizing the expected *penultimate* reward from the bandit that halts and ends the process. That is, if a bandit is activated and halts, stopping the control process, the controller receives the reward that bandit offered when it was activated, rather than the reward it halts on. Additionally in this section, we assume that the reward processes from each bandit are non-increasing. In fact, under this restriction, we may even maximize the reward *almost surely*. This result, while intuitive, acts as the workhorse for future optimality results.

We define the *penultimate solo payout* value of a policy π as,

$$\begin{aligned} V_{\pi}^{PSP}(\mathbb{X}) &= \mathbb{E}\left[X_{\pi}(\sigma_{\pi}-1)|\mathscr{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}\{i=\pi(\sigma_{\pi}-1)\}X^{i}_{T^{i}_{\pi}(\sigma_{\pi}-1)}|\mathscr{G}_0\right]. \end{aligned} \tag{4.8}$$

Theorem 11 (A Greedy Result for Non-Increasing Solo Payout Processes) *Given a collection of reward processes $\mathbb{X}$ such that for each i, X^i is almost surely non-increasing for $t < \sigma^i$, there exists a policy $\pi^* \in \mathscr{P}$ such that for any policy $\pi \in \mathscr{P}$,*

$$X_{\pi}(\sigma_{\pi}-1) \leq X_{\pi^*}(\sigma_{\pi^*}-1) \;\; (\mathbb{P}\text{-a.e.}). \tag{4.9}$$

In particular, such a π^* is given by the following greedy rule: In each round $s \geqslant 0$, activate the bandit with the largest current value of X^i, $\pi^*(s) = argmax_i\, X^i_{T^i_{\pi^*}(s)}$.

Proof. The proof proceeds by incremental improvements on an arbitrary policy.

Let $X^i_0 = \max_j X^j_0$. Let $\pi \in \mathscr{P}$ be arbitrary, and define $S = S^i_\pi(0)$, the first round bandit i is activated under π. If i is never activated, we take S to be infinite.

From π, we construct a policy $\pi' \in \mathscr{P}$ as follows: π' activates bandits in the same order as π, but it advances the first activation of bandit i from round $s = S$ to round $s = 0$. That is,

$$\pi'(s) = \begin{cases} i & \text{for } s = 0, \\ \pi(s-1) & \text{for } s = 1, 2, \dots S, \\ \pi(s) & \text{for } s \geqslant S+1. \end{cases} \tag{4.10}$$

It is important to observe that π' is in $\mathscr{P}$, as at every round s, the information available under π' is greater than or equal to the information available for the corresponding activation under π.

In the case that $\sigma_\pi > S+1$, that is π halts *after* the first activation of bandit i, then there is no difference between the rewards returned by either policy. Similarly, if $\sigma_\pi = S+1$, that is π halts *due to* the first activation of bandit i, the reward returned under π is X^i_0, and as bandit i halted on its first activation, the reward returned under π' is also X^i_0. In fact, it follows similarly that the only situation in which π and π' differ in their returned rewards is when $\sigma_\pi \leq S$ and $\sigma^i = 1$. Therefore,

$$\begin{aligned} X_{\pi'}(\sigma_{\pi'} - 1) - X_\pi(\sigma_\pi - 1) &= (X_{\pi'}(\sigma_{\pi'} - 1) - X_\pi(\sigma_\pi - 1))\mathbb{1}_{\{\sigma_\pi \leq S\}}\mathbb{1}_{\{\sigma^i = 1\}} \\ &= (X^i_0 - X_\pi(\sigma_\pi - 1))\mathbb{1}_{\{\sigma_\pi \leq S\}}\mathbb{1}_{\{\sigma^i = 1\}} \\ &\geqslant 0 \ \ (\mathbb{P}\text{-a.e.}). \end{aligned} \tag{4.11}$$

The last step follows taking X^i_0 as the initial largest reward, and that all bandits are non-increasing.

It follows that advancing the activation of the initial maximal bandit improves or at least does not change the value of a policy. This same argument can be applied at every round that follows, that at every step, activation of the current initial maximal bandit is an improvement over (or at least does not change the value) of any other policy. Note, collisions may occur if at a given round two bandits have equal rewards. This may be resolved at the discretion of the controller, such as by always taking the bandit with the smaller index i.

As each bandit halts in a finite time, almost surely, for sufficiently many greedy improvements as outlined above, the resulting improvement of any policy π will return the same value as the completely greedy strategy π^*. Hence,

$$X_\pi(\sigma_\pi - 1) \leq X_{\pi^*}(\sigma_{\pi^*} - 1) \ \ (\mathbb{P}\text{-a.e.}). \tag{4.12}$$

4.2.1 The Necessity of Finite Halting Times

This model is restricted in few ways, but one significant restriction is the assumption that $\sigma^i < \infty$ almost surely, for each bandit i. This assumption excludes cases such as the following, in which no optimal policy exists:

Consider two bandits, Bandit A offering a potential reward of \$100 in each time step, and Bandit B offering a potential reward of \$50 in each time step. Further, suppose that $\mathbb{P}^A(\sigma^A < \infty) = 0.5$, and $\sigma^B = 1$ almost surely - that is, Bandit B halts after its first activation.

Any policy on these bandits may be described in the following way: For $\tau \geqslant 0$ as a finite $\mathbb{F}^A$-stopping time, π_τ activates Bandit A until τ, then Bandit B, ending the process. The value of such a policy is given by

$$V_\tau^{PSP}(A,B) = \$100\ \mathbb{P}^A(\sigma^A < \tau) + \$50\ \mathbb{P}^A(\sigma^A \geqslant \tau) \leq 75. \tag{4.13}$$

This upper bound may be achieved within an arbitrary amount by choosing a finite, sufficiently large τ - the larger the τ, the closer to achieving the upper bound of \$75. However, taking τ to be infinite, the \$100 is only collected with probability 0.5, and Bandit B is never activated at all, yielding a total expected value of $\$100 \times 0.5 = \$50 < \$75$. In this case, there exist ε-optimal policies, but no optimal policy.

4.3 Maximizing Collective Payouts

In this section, we consider a model where rewards are collective, received from all bandits, at the final round of the process. We define the *collective payout* value of a policy π as,

$$V_\pi^{CP}(\mathbb{X}) = \sum_{i=1}^{N} \mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} | \mathscr{G}_0\right], \tag{4.14}$$

the expected total reward from all bandits at the end of the process. In the following subsections, we develop a policy $\pi^* \in \mathscr{P}$ such that for all $\pi \in \mathscr{P}$,

$$V^{CP}_{\pi}(\mathbb{X}) \leq V^{CP}_{\pi^*}(\mathbb{X}) \text{ } (\mathbb{P}\text{-a.e.}). \tag{4.15}$$

Remark 7. For algebraic convenience, we take $X^i_0 = 0$ for all i. For a more arbitrary reward processes $\{\hat{X}^i\}$, recall that the initial $\hat{X}^i_0$ are taken to be constant and known at the initial round by assumption. Hence, defining $X^i_t = \hat{X}^i_t - \hat{X}^i_0$, maximizing the total expected reward from the $\{\hat{X}^i\}$ processes is equivalent to maximizing the total expected reward from the $\{X^i\}$ processes.

4.3.1 Block Values

This section introduces a way of considering the 'value' of a set of activations of a bandit. The 'true' value of a decision to activate a bandit is not simply the reward gained through that decision, but instead must balance the immediate reward with the incurred probability of ending the control process through that decision, and the resulting loss of potential future rewards.

For each bandit i, for a given policy π we define τ^i_π to be the first activation of bandit i that *does not occur under* π. That is,

$$\tau^i_\pi = \min\{t \geqslant 0 : S^i_\pi(t) \geqslant \sigma_\pi\}. \tag{4.16}$$

With this, we state the following definitions:

Definition 2 (Process Blocks and their Values) *Given times $t' < t''$ with $t' < \sigma^i$, and a policy $\pi \in \mathscr{P}$ with $S^i_\pi(t') < \sigma_\pi$:*

1. *The* solo-payout value of the $[t', t'')$ - block of X^i *as:*

$$\rho^i(t', t'') = \frac{\mathbb{E}^i\left[X^i_{\sigma^i \wedge t''} - X^i_{t'} \middle| \mathscr{F}^i(t')\right]}{\mathbb{P}^i\left(t' < \sigma^i \leq t'' \middle| \mathscr{F}^i(t')\right)}. \tag{4.17}$$

2. *The* π-value of the $[t', t'')$ - block of X^i *as:*

$$v^i_\pi(t', t'') = \frac{\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi) \wedge t''} - X^i_{t'} \middle| \mathscr{H}^i_\pi(t')\right]}{\mathbb{P}\left(t' < \sigma^i \leq \tau^i_\pi \wedge t'' \middle| \mathscr{H}^i_\pi(t')\right)}. \tag{4.18}$$

The denominator of v^i_π may be interpreted as the probability that the process ends *due to bandit i*, halting during activation of the $[t', t'')$-block. Due to Eq. (4.2), the denominators of both block values are non-zero. The above quantities are all measurable with respect to the indicated σ-fields, and finite ($\mathbb{P}^i, \mathbb{P}$ -a.e.), due to (4.1).

Remark 8. The above might be justified as the 'value' of a block of activations in the following way: even if the incremental reward gained due to an activation block (the numerators) is small, if the probability of halting due to those activations (the denominators) is sufficiently small, there is very little risk in attempting to gain that increment through that activation. In fact, there might be more to gain in such a case than if the incremental reward were slightly larger, but the probability of halting were also larger. The above values captures this trade-off between risk of halting and reward gained.

Notionally, ρ^i can be thought of as the value of a block under sequential, consecutive activation, while v^i_π is, correspondingly, the value of a block potentially 'diluted' or broken up by activations of other bandits under π. The following theorem illustrates the relationship between ρ^i and v^i_π, essentially stating that the value of any block under some policy π is *at most* the value of *some* block activated consecutively.

Theorem 12 (Block Value Comparison) *For bandit i under policy π, for any time t_0 such that $S^i_\pi(t_0) < \sigma_\pi$, the following holds for any $\mathbb{H}^i_\pi$-stopping time τ with $t_0 < \tau$:*

$$v^i_\pi(t_0, \tau) \leq \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \rho^i(t_0, \hat{\tau}) \ (\mathbb{P}\text{-a.e.}). \tag{4.19}$$

Note that it follows from Eqs. (4.1, 4.2) that the essential supremum is finite *($\mathbb{P}$-a.e).*

Proof. For each bandit i and any $\pi \in \mathscr{P}$, it can be shown by cases (whether the control process does or does not end due to an activation of i) that $T^i_\pi(\sigma_\pi) = \sigma^i \wedge \tau^i_\pi$.

Therefore, for a given $\tau \in \hat{\mathbb{H}}^i_\pi(t_0)$,

$$\begin{aligned} \nu^i_\pi(t_0,\tau) &= \frac{\mathbb{E}\left[X^i_{\sigma^i\wedge\tau^i_\pi\wedge\tau} - X^i_{t_0}\middle|\mathscr{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \le \tau^i_\pi \wedge \tau \middle| \mathscr{H}^i_\pi(t_0)\right)} \\ &= \frac{\mathbb{E}\left[X^i_{\sigma^i\wedge(\tau^i_\pi\wedge\tau)} - X^i_{t_0}\middle|\mathscr{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \le (\tau^i_\pi \wedge \tau) \middle| \mathscr{H}^i_\pi(t_0)\right)} \\ &\le \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{H}}^i_\pi(t_0)} \frac{\mathbb{E}\left[X^i_{\sigma^i\wedge\hat{\tau}} - X^i_{t_0}\middle|\mathscr{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \le \hat{\tau} \middle| \mathscr{H}^i_\pi(t_0)\right)} \ (\mathbb{P}\text{-a.e.}). \end{aligned} \tag{4.20}$$

The last step above follows as, given that τ^i_π and τ are both in $\hat{\mathbb{H}}^i_\pi(t_0)$ by assumption, so too is $\tau^i_\pi \wedge \tau$.

Defining a 'global' π-analog of ρ^i,

$$\rho^i_\pi(t',t'') = \frac{\mathbb{E}\left[X^i_{\sigma^i\wedge t''} - X^i_{t'}\middle|\mathscr{H}^i_\pi(t')\right]}{\mathbb{P}\left(t' < \sigma^i \le t'' \middle| \mathscr{H}^i_\pi(t')\right)}, \tag{4.21}$$

we have the following relations:

$$\nu^i_\pi(t_0,\tau) \le \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{H}}^i_\pi(t_0)} \rho^i_\pi(t_0,\hat{\tau}) \le \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)} \rho^i(t_0,\hat{\tau}) \ (\mathbb{P}\text{-a.e.}). \tag{4.22}$$

The first is simply a restatement of Eq. (4.20). The second relation, the exchange from $\mathbb{H}^i_\pi$-stopping times to $\mathbb{F}^i$-stopping times, is intuitive: as the X^i process and σ^i are independent of the non-i bandits, information about those independent bandits (through the $\mathbb{H}^i_\pi$-stopping times) cannot assist in maximizing the quotient. Rigorously, this amounts to integrating out the independent bandits, and is done in more detail in Section 4.6 as Proposition 10.

The following proposition provides, using ρ^i and ν^i_π, alternative expressions for the incremental reward gained through the activation of a block.

Proposition 4 *For each bandit i, the following hold for any $\mathbb{F}^i$-stopping times $\tau' < \tau''$ where the quantities are defined. Equality also holds when conditioning with respect to the initial information, $\mathscr{F}^i(0)$, $\mathscr{G}_0$ respectively via the tower property.*

$$\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau''} - X^i_{\tau'}\middle|\mathscr{F}^i(\tau')\right] = \mathbb{E}^i\left[\sum_{t=\tau'}^{\tau''-1} \rho^i(\tau',\tau'')\mathbb{1}_{\{\sigma^i=t+1\}}\middle|\mathscr{F}^i(\tau')\right] \tag{4.23}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)\wedge\tau''} - X^i_{\tau'}\middle|\mathscr{H}^i_\pi(\tau')\right] = \mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1} \nu^i_\pi(\tau',\tau'')\mathbb{1}_{\{\sigma_\pi=S^i_\pi(t)+1\}}\middle|\mathscr{H}^i_\pi(\tau')\right] \tag{4.24}$$

Proof. The above equations follow directly from Eqs. (4.17, 4.18), observing the following relations:

$$\begin{aligned}\mathbb{P}^i\left(t' < \sigma^i \leq t'' \middle| \mathscr{F}^i(t')\right) &= \mathbb{E}^i\left[\sum_{t=t'}^{t''-1} \mathbb{1}_{\{\sigma^i = t+1\}} \middle| \mathscr{F}^i(t')\right], \\ \mathbb{P}\left(t' < \sigma^i \leq \tau^i_\pi \wedge t'' \middle| \mathscr{H}^i_\pi(t')\right) &= \mathbb{E}\left[\sum_{t=t'}^{t''-1} \mathbb{1}_{\{\sigma_\pi = S^i_\pi(t)+1\}} \middle| \mathscr{H}^i_\pi(t')\right].\end{aligned} \tag{4.25}$$

4.3.2 Solo Payout Indices and Times

Theorem 12 indicates the significance of the following quantity.

Definition 3 (The Solo-Payout Index) *For any $t < \sigma^i$, the incremental* Solo-Payout Index at t *is defined to be*

$$\rho^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \rho^i(t, \tau). \tag{4.26}$$

This index, interpreted as the maximal quotient of 'incremental reward' over 'probability of termination/halting' as in Eq. (4.17) was anticipated by Sonin in [66], who defined it over Markov chain reward processes as a generalization of the Gittins index.

The following result demonstrates that $\rho^i(t)$ is realized as the value of *some* block from time t, that is for some $\tau > t$, $\rho^i(t) = \rho^i(t, \tau)$ ($\mathbb{P}^i$-a.e.). As such, $\rho^i(t)$ represents the *maximal block value achievable from process i from time t.*

Proposition 5 *For any time $t_0 < \sigma^i$, there exists a $\tau \in \hat{\mathbb{F}}^i(t_0)$ such that $\rho^i(t_0) = \rho^i(t_0, \tau)$* ($\mathbb{P}^i$-a.e.).

The proof is somewhat technical and not the focus, and hence is relegated to Section 4.6.

The solo-payout indices and their realizing blocks provide a natural time scale with which to view a process, in terms of a sequence of blocks. In particular, we define the following sequence:

Definition 4 (Solo-Payout Index Times) *Define a sequence of $\mathbb{F}^i$-stopping times $\{\tau^i_k\}_{k \geqslant 0}$ in the following way, that $\tau^i_0 = 0$, and for $k > 0$,*

$$\tau^i_{k+1} = \arg\operatorname{ess\,sup}\{\rho^i(\tau^i_k, \tau) : \tau \in \hat{\mathbb{F}}^i(\tau^i_k)\}. \tag{4.27}$$

In the case that $\tau^i_k = \sigma^i$ for some k, then $\tau^i_{k'}$ is taken to be infinite for all larger k'. In the case that $\tau^i_k < \sigma^i$, we have that $\rho^i(\tau^i_k) = \rho^i(\tau^i_k, \tau^i_{k+1})$. The question of whether the 'arg ess sup' exists is resolved in the positive by Proposition 5; if there is more than one stopping time that attains the 'arg ess sup', we take τ^i_{k+1} to be the one demonstrated by the application of Lemma 9 in the proof of Proposition 5.

Using this sequence of stopping times, we partition the local process times $\mathbb{N}^i = \{0,1,2,\ldots\}$ into

$$\mathbb{N}^i = [0, \tau^i_1) \cup [\tau^i_1, \tau^i_2) \cup [\tau^i_2, \tau^i_3) \cup \ldots.$$

One important property of this partition is the following:

Proposition 6 (Solo-Payout Indices Non-Increasing over Index Times) *For any $k > 0$ such that $\tau^i_k < \sigma^i$, the following is true: $\rho^i(\tau^i_{k-1}) \geqslant \rho^i(\tau^i_k)$* ($\mathbb{P}^i$-a.e.).

For intuition, recall the $\{\tau^i_k\}_k$ are meant to realize successively the maximal indices of the process $\{X^i_t\}_t$. If $\rho^i(\tau^i_{k-1}) = \rho^i(\tau^i_{k-1}, \tau^i_k) < \rho^i(\tau^i_k)$, the index from τ^i_{k-1} may be increased by taking a block that extends from τ^i_{k-1} *past* τ^i_k. This contradicts the idea of the $\{\tau^i_k\}_k$ as realizing the maximal indices. The proof is relegated to the Section 4.6 as technical, and not the focus of this work.

4.3.3 Equivalent Solo Payout Processes

For each bandit, we have developed a partition of local time into blocks of activations via the solo payout index stopping times. With Proposition 4 in mind, we use these blocks to define a set of reward equivalent solo payout processes, and π-equivalent solo payout processes.

Definition 5 *Given the collection of* reward processes $\mathbb{X} = (X^1, ..., X^N)$*, and $\{\tau^i_k\}_{k \geqslant 0}$ for each i as in Definition 4, we define:*

1. *The* reward-equivalent solo payout collection $\mathbb{Y}^X = (Y^1, ..., Y^N)$ *by*

$$Y^i(t) = \rho^i(\tau^i_k), \quad \textit{if } \tau^i_k \leq t < \tau^i_{k+1}. \tag{4.28}$$

2. *For* $\pi \in \mathscr{P}$*, the* π-equivalent solo payout collection $\mathbb{Y}^X_\pi = (Y^1_\pi, ..., Y^N_\pi)$*, by*

$$Y^i_\pi(t) = \nu^i_\pi(\tau^i_k, \tau^i_{k+1}), \quad \textit{if } \tau^i_k \leq t < \tau^i_{k+1}. \tag{4.29}$$

Like X^i, the process Y^i is defined on $(\Omega^i, \mathscr{F}^i, \mathbb{P}^i, \mathbb{F}^i)$ and is $\mathbb{F}^i$-adapted, as the $\rho^i(\tau^i_k)$ is defined by the information available locally at time τ^i_k. However, as the $\nu^i_\pi(\tau^i_k, \tau^i_{k+1})$ depend on the specifics of policy π, so do the Y^i_π processes; the Y^i_π processes are $\mathbb{H}^i_\pi$-adapted, but not $\mathbb{F}^i$-adapted. Note, Y^i is only really defined for $t < \sigma^i$, and Y^i_π is only defined for t such that $S^i_\pi(t) < \sigma_\pi$. However, since no rewards are collected from bandit i after these times, this lack of definition is of no concern.

The following are simple, but important properties of the $\mathbb{Y}^X, \mathbb{Y}^X_\pi$ processes.

Proposition 7 *For* $\pi \in \mathscr{P}$*, for each i, and any k where the following quantities are defined,*

$$\mathbb{E}^i\left[X^i_{\sigma^i \wedge \tau^i_{k+1}} - X^i_{\tau^i_k} \middle| \mathscr{F}^i(\tau^i_k)\right] = \mathbb{E}^i\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t) \mathbb{1}_{\{\sigma^i = t+1\}} \middle| \mathscr{F}^i(\tau^i_k)\right], \tag{4.30}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi) \wedge \tau^i_{k+1}} - X^i_{\tau^i_k} \middle| \mathscr{H}^i_\pi(\tau^i_k)\right] = \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i_\pi(t) \mathbb{1}_{\{\sigma_\pi = S^i_\pi(t)+1\}} \middle| \mathscr{H}^i_\pi(\tau^i_k)\right]. \tag{4.31}$$

As with Proposition 4, equality also holds when conditioning with respect to $\mathscr{F}^i(0), \mathscr{G}_0$.

Proof. *This follows as an application of Proposition 4 and the definitions of* Y^i, Y^i_π.

The following proposition serves as justification of the term 'equivalent' in describing the $\mathbb{Y}^X, \mathbb{Y}^X_\pi$ collections.

Proposition 8 *For each i, for any policy* $\pi \in \mathscr{P}$*,*

$$\mathbb{E}^i\left[X^i_{\sigma^i} \middle| \mathscr{F}^i(0)\right] = \mathbb{E}^i\left[Y^i(\sigma^i - 1) \middle| \mathscr{F}^i(0)\right], \tag{4.32}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} \middle| \mathscr{G}_0\right] = \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_\pi - 1)\}} Y^i_\pi(T^i_\pi(\sigma_\pi - 1)) \middle| \mathscr{G}_0\right]. \tag{4.33}$$

Proof. *Each follows from the corresponding equation in Prop. 7, summing over k and taking expectations from the initial time, via the tower property. On the right hand sides, the* X^i *terms telescope in the sum, and* X^i_0 *is taken to be 0. On the left hand sides, the sums over Y may be* *expressed as single terms, due to the indicators.*

Proposition 9 *For each i, for each time $t > 0$ such that $Y^i(t)$ is defined,*

$$Y^i(t-1) \geqslant Y^i(t) \text{ (}\mathbb{P}^i\text{-a.e.) .} \tag{4.34}$$

Proof. *This follows immediately from Proposition 6, and Definition 5.1.*

Theorem 13 (Comparison of Equivalent, π-Equivalent Solo Payout Processes) *For any $\pi \in \mathscr{P}$, for each i and all time t where both are defined, we have:*

$$Y^i_\pi(t) \leq Y^i(t) \quad (\mathbb{P}\text{-a.e.}). \tag{4.35}$$

Proof. *For such a t, we have for some k that $\tau^i_k \leq t < \tau^i_{k+1}$, and as an application of Theorem 12,*

$$\begin{aligned} Y^i_\pi(t) = v^i_\pi(\tau^i_k, \tau^i_{k+1}) &\leq \operatorname*{ess\,sup}_{\tau' \in \hat{\mathbb{H}}^i_\pi(\tau^i_k)} v^i_\pi(\tau^i_k, \tau') \\ &\leq \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(\tau^i_k)} \rho^i(\tau^i_k, \hat{\tau}) = \rho^i(\tau^i_k) = Y^i(t) \quad (\mathbb{P}\text{-a.e.}). \end{aligned} \tag{4.36}$$

4.3.4 The Optimal Policy for Collective Payout Bandits

The derivation of the optimal control policy for an arbitrary collection of reward processes $\mathbb{X}$ under a collective reward structure is all but immediate now.

Theorem 14 (The Optimal Collective Payout Control Policy) *For a collection of reward processes $\mathbb{X} = (X^1, X^2, \ldots, X^N)$, and the associated stopping times $\{\sigma^i\}_{i=1,\ldots,N}$, there exists a strategy $\pi^* \in \mathscr{P}$ such that for all $\pi \in \mathscr{P}$,*

$$V^{CP}_\pi(\mathbb{X}) \leq V^{CP}_{\pi^*}(\mathbb{X}) \text{ (}\mathbb{P}\text{-a.e.).} \tag{4.37}$$

In particular, such an optimal policy π^* can be described in the following way: successively activate the bandit with the largest current solo payout index,

$$\rho^i(t) = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau} - X^i_t \middle| \mathscr{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathscr{F}^i(t)\right)}, \tag{4.38}$$

for the duration of the corresponding index block.

Before giving the proof of this theorem, we give a corollary, which gives a useful alternative characterization of the policy π^*.

Corollary 1 *An alternative characterization of the policy π^* in Theorem 14 is the following: at every round, activate the bandit with the largest current solo payout index.*

Proof. From Theorem 14, it follows that the optimal first activation is to activate a bandit with the largest current solo payout index. If that activation does not halt the bandit and end the process, the controller is faced with a structurally identical decision problem. It follows that again, the optimal action is to activate a bandit with the largest current solo payout index. This argument may be iterated until halting, which will occur in finite time by assumption on the $\{\sigma^i\}$.

Proof. [of Theorem 14.] For an arbitrary policy π, and π^* as indicated above, we establish the following relations:

$$V^{CP}_\pi(\mathbb{X}) = V^{PSP}_\pi(\mathbb{Y}^X_\pi) \leq V^{PSP}_\pi(\mathbb{Y}^X) \leq V^{PSP}_{\pi^*}(\mathbb{Y}^X) = V^{CP}_{\pi^*}(\mathbb{X}) \text{ } (\mathbb{P}\text{-a.e.}), \tag{4.39}$$

i.e., for any policy π, we have that $V^{CP}_\pi(\mathbb{X}) \leq V^{CP}_{\pi^*}(\mathbb{X})$ ($\mathbb{P}$-a.e.) and therefore π^* is an optimal policy.

In the following steps we prove relations (4.39).

Step 1: $V^{CP}_\pi(\mathbb{X}) = V^{PSP}_\pi(\mathbb{Y}^X_\pi)$, ($\mathbb{P}$-a.e.).

We have, via Prop. 8, Eq. (4.33),

$$\begin{aligned} V^{CP}_\pi(\mathbb{X}) &= \sum_{i=1}^{N} \mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} \middle| \mathscr{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}} Y^i_\pi(T^i_\pi(\sigma_\pi - 1)) \middle| \mathscr{G}_0\right] = V^{PSP}_\pi(\mathbb{Y}^X_\pi). \end{aligned} \tag{4.40}$$

Note, because the Y^i_π processes are defined in terms of π, they are not $\mathbb{F}^i$-adapted, and cannot be utilized under any other policy. However, the value $V^{PSP}_\pi(\mathbb{Y}^X_\pi)$ is well defined via the above equation.

Step 2: $V^{PSP}_\pi(\mathbb{Y}^X_\pi) \leq V^{PSP}_\pi(\mathbb{Y}^X)$ ($\mathbb{P}$-a.e.).

This follows from the point-wise inequality of Theorem 13, $Y^i_\pi(t) \leq Y^i(t)$ for all t. Note that for any t where $Y^i_\pi(t)$ is not defined, the t^{th} activation of i does not occur under π, and no comparison is necessary.

Step 3: $V^{PSP}_\pi(\mathbb{Y}^X) \leq V^{PSP}_{\pi^*}(\mathbb{Y}^X)$ ($\mathbb{P}$-a.e.).

This follows simply from Theorem 11: by construction, the terms of each Y^i process are equal to the solo payout indices of X^i, piecewise constant over blocks, and non-increasing.

Step 4: $V^{PSP}_{\pi^*}(\mathbb{Y}^X) = V^{CP}_{\pi^*}(\mathbb{X})$ ($\mathbb{P}$-a.e.).

Note that π^* activates bandits consecutively over the duration of their index blocks. For a given i, define

$$k_i^* = \min_{k \geqslant 0} \{ S^i_{\pi^*}(\tau^i_k) \geqslant \sigma_\pi \}, \tag{4.41}$$

the first block of i that is *not* activated under π^*. Note then that for each i, we have the following relation

$$T^i_{\pi^*}(\sigma_{\pi^*}) = \sigma^i \wedge \tau^i_{k_i^*}. \tag{4.42}$$

Expressing the value of policy π^* relative to activations over blocks, and utilizing the tower property, we have the following equivalences:

$$\begin{aligned} V^{PSP}_{\pi^*}(\mathbb{Y}^X) &= \sum_{i=1}^N \sum_{k=0}^\infty \mathbb{E}\left[\mathbb{1}_{\{k_i^* > k\}} \sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t) \mathbb{1}_{\{\sigma^i = t+1\}} \big| \mathscr{G}_0 \right] \\ &= \sum_{i=1}^N \sum_{k=0}^\infty \mathbb{E}\left[\mathbb{1}_{\{k_i^* > k\}} \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t) \mathbb{1}_{\{\sigma^i = t+1\}} \big| \mathscr{H}^i_\pi(\tau^i_k) \right] \big| \mathscr{G}_0 \right] \\ &= \sum_{i=1}^N \sum_{k=0}^\infty \mathbb{E}\left[\mathbb{1}_{\{k_i^* > k\}} \mathbb{E}\left[X^i_{\sigma^i \wedge \tau^i_{k+1}} - X^i_{\tau^i_k} \big| \mathscr{H}^i_\pi(\tau^i_k) \right] \big| \mathscr{G}_0 \right] \\ &= \sum_{i=1}^N \mathbb{E}\left[X^i_{\sigma^i \wedge \tau^i_{k_i^*}} - X^i_0 \big| \mathscr{G}_0 \right] \\ &= \sum_{i=1}^N \mathbb{E}\left[X^i_{T^i_{\pi^*}(\sigma_{\pi^*})} \big| \mathscr{G}_0 \right] = V^{CP}_{\pi^*}(\mathbb{X}). \end{aligned} \tag{4.43}$$

Note the exchange over blocks of the Y^i rewards for the X^i rewards is due to Proposition 7, Eq. (4.30), taking the extension to $\mathscr{H}^i_{\pi^*}(\tau^i_k)$ in place of $\mathscr{F}^i(\tau^i_k)$.

Remark 9. The above theorem demonstrates a policy $\pi^* \in \mathscr{P}$ that is $\mathbb{P}$-a.e. superior (or equivalent) to every other policy $\pi \in \mathscr{P}$. However, the set of non-anticipatory policies $\mathscr{P}$ was defined in a fairly restrictive sense in Section 4.1.1, so that the decision in any round was completely determined by the results of the past. This might be weakened to allow for randomized policies, so that the decision in a given round might depend on the results of independent events, e.g., coin flips. However, such a construction simply amounts to placing a distribution on $\mathscr{P}$. Since π^* is $\mathbb{P}$-a.e. superior to any $\pi \in \mathscr{P}$, π^* would be similarly superior to any policy sampled randomly from $\mathscr{P}$.

4.4 Alternative Payout Schemes

Utilizing the results of the previous section, we may provide index policies for optimizing the rewards/costs from a number of alternative payout models, by reducing them to the model of the previous section.

4.4.1 Maximizing Solo Payouts

In this section, we consider the problem of maximizing the final reward from the bandit that halts the process. We define the *solo payout* value of a policy π as,

$$\begin{aligned} V^{SP}_{\pi}(\mathbb{X}) &= \mathbb{E}\left[X_{\pi}(\sigma_{\pi})|\mathscr{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_{\pi}-1)\}} X^i_{T^i_{\pi}(\sigma_{\pi})} |\mathscr{G}_0\right]. \end{aligned} \tag{4.44}$$

To provide an index policy to maximize this value function, we reduce it to the previous model in the following way: define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1\leq i\leq N}$ by for each i, each $t \geqslant 0$,

$$Z^i_t = \mathbb{1}_{\{\sigma^i = t\}} X^i_t. \tag{4.45}$$

Notice that at round σ_{π}, $Z^i_t = 0$ for all bandits that did not halt the process, and $Z^i_t = X^i_{\sigma^i}$ for the bandit that did halt the process. Hence the collective payout under $\mathbb{Z}$ is equal to the solo payout under $\mathbb{X}$, $V^{CP}_{\pi}(\mathbb{Z}) = V^{SP}_{\pi}(\mathbb{X})$. Applying the previous results in this case, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the solo payout under $\mathbb{X}$, given by a policy that always activates bandits according to the maximum *solo payout index*:

$$\rho^i_{SP}(t) = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\tau\geqslant\sigma^i\}}X^i_{\sigma^i}\middle|\mathscr{F}^i(t)\right]}{\mathbb{P}^i\left(t<\sigma^i\leq\tau\middle|\mathscr{F}^i(t)\right)}. \tag{4.46}$$

It is interesting to observe that the policy based on the above index has a very natural interpretation: viewing the index as the maximal conditional expected payout of a bandit on its halting, the policy always activates the bandit with the largest potential payout - should it pay out.

4.4.2 Minimizing Non-Halting Cost

In this section, we consider the case in which the controller *pays a cost* based on the bandits that did not halt the process, and wishes to minimize this cost. We define the *halting cost* of a policy π as:

$$\begin{aligned} C^H_{\pi}(\mathbb{X}) &= \mathbb{E}\left[\sum_{i\neq\pi(\sigma_\pi-1)} X^i_{T^i_\pi(\sigma_\pi)}|\mathscr{G}_0\right] \\ &= \sum_{i=1}^{N}\mathbb{E}\left[\mathbb{1}_{\{i\neq\pi(\sigma_\pi-1)\}}X^i_{T^i_\pi(\sigma_\pi)}|\mathscr{G}_0\right]. \end{aligned} \tag{4.47}$$

To provide an index policy to minimize this value function, we reduce it to a previous model in the following way: define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1\leq i\leq N}$ by for each i, each $t \geqslant 0$,

$$Z^i_t = -\mathbb{1}_{\{\sigma^i\neq t\}}X^i_t. \tag{4.48}$$

Notice that at round σ_π, if bandit i was activated to halt the process ($\pi(\sigma_\pi - 1) = i$), $Z^i_t = 0$, otherwise $Z^i_t = -X^i_t$. Hence, the collective payout under $\mathbb{Z}$ is equal to the negative of the halting cost under $\mathbb{X}$, $V^{CP}_{\pi}(\mathbb{Z}) = -C^H_{\pi}(\mathbb{X})$. Maximizing the collective payout under $\mathbb{Z}$ therefore minimizes the halting cost under $\mathbb{X}$. Applying the results of the previous section, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the halting cost under $\mathbb{X}$, given by a policy that always activates bandits according to the maximum *halting cost index*:

$$\rho^i_{HC}(t) = -\operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\sigma^i>\tau\}}X^i_\tau - X^i_t\middle|\mathscr{F}^i(t)\right]}{\mathbb{P}^i\left(t<\sigma^i\leq\tau\middle|\mathscr{F}^i(t)\right)}. \tag{4.49}$$

4.4.3 Maximizing Collective Profit

We may combine the results of the previous two subsections in the following way: to each bandit i we associate a reward process $\{R^i_t\}_{t\geqslant 0}$ and a cost process $\{C^i_t\}_{t\geqslant 0}$. In this section, we consider the case where the controller gains a reward from the bandit that halts the process, and pays a cost for each bandit that does not halt. The controller wishes to maximize her total profit. We define the *collective profit* of a policy π as,

$$V^{PC}_\pi(\mathbb{R},\mathbb{C}) = \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}} R^i_{T^i_\pi(\sigma_\pi)} - \mathbb{1}_{\{i\neq\pi(\sigma_\pi-1)\}} C^i_{T^i_\pi(\sigma_\pi)} |\mathscr{G}_0\right]. \tag{4.50}$$

To provide an index policy to maximize this value function, we reduce it to a previous model in the following way: define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1\leq i\leq N}$ by for each i, each $t \geqslant 0$,

$$Z^i_t = \mathbb{1}_{\{\sigma^i=t\}} R^i_t - \mathbb{1}_{\{\sigma^i\neq t\}} C^i_t. \tag{4.51}$$

Notice that at round σ_π, $Z^i_t = -C^i_t$ for all bandits that did not halt the process, and $Z^i_t = R^i_t$ for the bandit that did halt the process. Hence the collective payout under $\mathbb{Z}$ is equal to the collective profit under $(\mathbb{R},\mathbb{C})$, $V^{CP}_\pi(\mathbb{Z}) = V^{PC}_\pi(\mathbb{R},\mathbb{C})$. Applying the previous results in this case, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the collective profit under $(\mathbb{R},\mathbb{C})$, given by a policy that always activates bandits according to the maximum *collective profit index*:

$$\rho^i_{PC}(t) = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\sigma^i\leq\tau\}} R^i_{\sigma^i} - \mathbb{1}_{\{\sigma^i>\tau\}} C^i_\tau + C^i_t \middle| \mathscr{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathscr{F}^i(t)\right)}. \tag{4.52}$$

4.5 Cumulative Payouts and Recovering the Gittins Index

In this section, we consider the case in which the controller gains a bandit's current reward each time that bandit is chosen to be activated. Bandits that are never activated give no rewards. The controller wishes to maximize her total payout. We define the *cumulative collective payout* value of a policy π as,

$$V^{CCP}_\pi(\mathbb{X}) = \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=0}^{T^i_\pi(\sigma_\pi)-1} X^i_t \middle| \mathscr{G}_0\right]. \tag{4.53}$$

Note, in the above expression we take empty sums, i.e., when $T^i_\pi(\sigma_\pi) = 0$, to be 0. To provide an index policy to maximize this value function, we reduce it to the previous model in the following way: define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1 \leq i \leq N}$ by, for each i, each $t \geqslant 0$,

$$Z^i_t = \sum_{t'=0}^{t-1} X^i_{t'}. \tag{4.54}$$

It follows easily that the collective payout under $\mathbb{Z}$ is equal to the collective cumulative payout under $\mathbb{X}$, $V^{CP}_\pi(\mathbb{Z}) = V^{CCP}_\pi(\mathbb{X})$. Applying the previous results in this case, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the collective cumulative payout under $\mathbb{X}$, given by a policy that always activates bandits according to the maximum *collective cumulative payout index*:

$$\rho^i_{CCP}(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\sum_{t'=t}^{\sigma^i \wedge \tau - 1} X^i_{t'} \middle| \mathscr{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathscr{F}^i(t)\right)}. \tag{4.55}$$

This extension of the collective payout model is not necessarily interesting in its own right, but consider the following: each time the controller activates a bandit, all potential future rewards are effectively reduced or discounted by a factor equal to the probability of that decision halting the process. In the special case that each halting time $\sigma^i > 0$ is a Geometric random variable with a constant parameter $0 < \beta < 1$, i.e., probability of survival, independent of the reward processes $\mathbb{X}$, this results in every activation discounting all future rewards by a factor of β. It follows that

$$V^{CCP}_\pi(\mathbb{X}) = \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^{T^i_\pi(\sigma_\pi)-1} X^i_t \middle| \mathscr{G}_0\right] = \mathbb{E}\left[\sum_{s=0}^{\infty} \beta^s X_\pi(s) \middle| \mathscr{G}_0\right]. \tag{4.56}$$

Maximizing the collective cumulative payout under this model is then equivalent to maximizing the total expected discounted reward of $\mathbb{X}$ under constant discount factor β, precisely the framework outlined by Gittins in the Markov case [25]. In this case, the collective cumulative payout index reduces to

$$\rho^i_{CCP}(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\sum_{t'=t}^{\tau-1} \beta^{t'-t} X^i_{t'} \middle| \mathscr{F}^i(t)\right]}{\mathbb{E}^i\left[1 - \beta^{\tau - t} \middle| \mathscr{F}^i(t)\right]} = \frac{1}{1-\beta} \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\sum_{t'=t}^{\tau-1} \beta^{t'} X^i_{t'} \middle| \mathscr{F}^i(t)\right]}{\mathbb{E}^i\left[\sum_{t'=t}^{\tau-1} \beta^{t'} \middle| \mathscr{F}^i(t)\right]}, \tag{4.57}$$

where the essential supremum on the right is precisely the Gittins index for bandit i. As $1/(1-\beta)$ is a constant, positive factor, activating according to the maximal collective cumulative payout index and activating according to the maximal Gittins index result in equivalent, optimal policies.

We may extend this equivalence of discounting and probabilities of survival in the following way: although in the Geometric case as above, β essentially represents the probability of survival / non-halting for a given bandit activation, we may consider an arbitrary, bandit-specific discounting sequence $\{\beta_0^i, \beta_1^i, \ldots\}$. We may interpret β_t^i as the probability of bandit *i not* halting on its t^{th} activation, or the depreciation incurred on rewards due to the t^{th} activation of bandit i. In this second case, if the $\{\beta_t^i\}_{t\geqslant 0}$ are taken to be unequal, this may be interpreted as the activations of a given bandit having potentially non-uniform durations. This model of non-uniform activation duration, as a generalization of the Gittins formulation, was considered in [15], and served as much of the inspiration for this present work.

4.6 Proofs

Proposition 10 *For bandit i under policy π, for any time t_0 such that $S_\pi^i(t_0) < \sigma_\pi$, the following holds:*

$$\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{H}}_\pi^i(t_0)} \rho_\pi^i(t_0, \hat{\tau}) \leq \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)} \rho^i(t_0, \hat{\tau}) \text{ ($\mathbb{P}$-a.e.)}. \tag{4.58}$$

Proof. Without loss of generality, we may take $t_0 = 0$. Recall the definition of ρ_π^i, ρ^i:

$$\begin{aligned} \rho_\pi^i(t', t'') &= \frac{\mathbb{E}\left[X_{\sigma^i\wedge t''}^i - X_{t'}^i \middle| \mathscr{H}_\pi^i(t')\right]}{\mathbb{P}\left(t' < \sigma^i \leq t'' \middle| \mathscr{H}_\pi^i(t')\right)}, \\ \rho^i(t', t'') &= \frac{\mathbb{E}^i\left[X_{\sigma^i\wedge t''}^i - X_{t'}^i \middle| \mathscr{F}^i(t')\right]}{\mathbb{P}^i\left(t' < \sigma^i \leq t'' \middle| \mathscr{F}^i(t')\right)}. \end{aligned} \tag{4.59}$$

Letting R denote the R.H.S. of Eq. (4.58), observe (by the definition of the essential supremum) that for any $\hat{\tau} \in \hat{\mathbb{F}}^i(0)$,

$$\mathbb{E}\left[X_{\sigma^i\wedge\hat{\tau}}^i - X_0^i - R\mathbb{1}\{0 < \sigma^i \leq \hat{\tau}\} \middle| \mathscr{F}^i(0)\right] \leq 0 \text{ ($\mathbb{P}$-a.e.)}. \tag{4.60}$$

To prove the proposition, it suffices to show that for any $\hat{\tau} \in \hat{\mathbb{H}}_\pi^i(0)$,

$$\mathbb{E}\left[X_{\sigma^i\wedge\hat{\tau}}^i - X_0^i - R\mathbb{1}\{0 < \sigma^i \leq \hat{\tau}\} \middle| \mathscr{H}_\pi^i(0)\right] \leq 0 \text{ ($\mathbb{P}$-a.e.)}. \tag{4.61}$$

For compactness of argument, we take $N = 2$ and $i = 1$, though the following argument generalizes to arbitrary bandits in the obvious way. For notational compactness, we define $W_t^i = X_{\sigma^i\wedge t}^i - X_0^i - R\mathbb{1}\{0 < \sigma^i \leq t\}$.

Note that for any set $A \in \mathscr{H}_\pi^1(0)$, and any $\tau \in \hat{\mathbb{H}}_\pi^1(0)$,

$$\mathbb{E}\left[\mathbb{1}_A \mathbb{E}\left[W_\tau^1 \middle| \mathscr{H}_\pi^1(0)\right]\right] = \mathbb{E}\left[\mathbb{1}_A W_\tau^1\right]. \tag{4.62}$$

Taking A as a rectangle in $\mathscr{H}_{\pi}^{1}(0)$, $A = A_1 \times A_2$, observe that $A_1 \in \mathscr{F}^1(0)$. The indicator may be decomposed as $\mathbb{1}_A(\omega) = \mathbb{1}_{A_1}(\omega^1)\mathbb{1}_{A_2}(\omega^2)$. It follows as a result of the initial integrability assumptions on the bandits, Eqs. (4.1), (4.2), that we may exchange the expectation over the product space for an iterated expectation:

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A W_\tau^1\right] &= \mathbb{E}^2\left[\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{1}_{A_2}W_\tau^1\right]\right]\\ &= \mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}W_\tau^1\right]\right]\\ &= \mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{E}^1\left[W_\tau^1\middle|\mathscr{F}^1(0)\right]\right]\right].\end{aligned} \tag{4.63}$$

Observe that, while τ (begin an $\mathbb{H}_{\pi}^{1}$-stopping time) may have a dependence on Ω^2, inside the iterated integral with the dependence on Ω^2 fixed, it is an $\mathbb{F}^i$-stopping time. Hence, as an application of Eq. (4.60), we have the bound

$$\mathbb{E}\left[\mathbb{1}_A W_\tau^1\right] = \mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{E}^1\left[W_\tau^1\middle|\mathscr{F}^1(0)\right]\right]\right] \leq \mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}0\right]\right] = 0. \tag{4.64}$$

Hence, for all rectangles $A \in \mathscr{H}_{\pi}^{1}(0)$, $\mathbb{E}\left[\mathbb{1}_A\mathbb{E}\left[W_\tau^1\middle|\mathscr{H}_{\pi}^{1}(0)\right]\right] \leq 0$. This extends via the usual monotone-class type argument to *all* $A \in \mathscr{H}_{\pi}^{1}(0)$. Hence, it follows that for all $\tau \in \hat{\mathbb{H}}_{\pi}^{1}(0)$,

$$\mathbb{E}\left[W_\tau^1\middle|\mathscr{H}_{\pi}^{1}(0)\right] \leq 0 \text{ } (\mathbb{P}\text{-a.e.}). \tag{4.65}$$

This establishes the result.

Proof. [of Proposition 5.]

The proof of Proposition 5 requires the following technical lemma, which follows from/in parallel with Proposition VI-1-3 in [65].

Lemma 9 *In an arbitrary probability space with a filtration $\mathbb{J} = \{\mathscr{J}_t\}_{t\geqslant 0}$, consider an adapted discrete-time process $\{Z_t\}_{t\geqslant 0}$ such that $\mathbb{E}\left[\sup_{\mathbb{N}}|Z_t|\middle|\mathscr{J}_0\right] < \infty$. If the $\mathbb{J}$-stopping time $\tau^* \in \hat{\mathbb{J}}(0)$ defined by*

$$\tau^* = \inf\{n > 0 : \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(n)} \mathbb{E}\left[Z_\tau\middle|\mathscr{J}_n\right] \leq Z_n\} \tag{4.66}$$

is almost surely finite, then

$$\mathbb{E}\left[Z_{\tau^*}\middle|\mathscr{J}_0\right] = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0)} \mathbb{E}\left[Z_\tau\middle|\mathscr{J}_0\right] \text{ } (\mathbb{P}\text{-a.e.}). \tag{4.67}$$

We have that for all $\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)$, $\rho^i(t_0, \hat{\tau}) \leq \rho^i(t_0)$ ($\mathbb{P}^i$-a.e.). Taking

$$\mathbb{P}^i(t_0 < \sigma^i \leq \hat{\tau} \big| \mathscr{F}^i(t_0)) = \mathbb{E}^i \left[\mathbb{1}_{\{t_0 < \sigma^i \leq \hat{\tau}\}} \big| \mathscr{F}^i(t_0) \right],$$

we have in parallel with Eq. (4.23),

$$\mathbb{E}^i \left[X^i_{\sigma^i \wedge \hat{\tau}} - X^i_{t_0} - \rho^i(t_0) \mathbb{1}_{\{t_0 < \sigma^i \leq \hat{\tau}\}} \big| \mathscr{F}^i(t_0) \right] \leq 0 \; (\mathbb{P}^i\text{-a.e.}). \tag{4.68}$$

Defining

$$\varepsilon = - \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i \left[X^i_{\sigma^i \wedge \hat{\tau}} - X^i_{t_0} - \rho^i(t_0) \mathbb{1}_{\{t_0 < \sigma^i \leq \hat{\tau}\}} \big| \mathscr{F}^i(t_0) \right], \tag{4.69}$$

we have that $\varepsilon \geqslant 0$ ($\mathbb{P}^i$-a.e.). We may use $-\varepsilon$ as an improved upper bound in Eq. (4.68). This may be rearranged to yield

$$\rho^i(t_0, \hat{\tau}) \leq \rho^i(t_0) - \frac{\varepsilon}{\mathbb{E}^i \left[\mathbb{1}_{\{t_0 < \sigma^i \leq \hat{\tau}\}} \big| \mathscr{F}^i(t_0) \right]} \leq \rho^i(t_0) - \varepsilon \; (\mathbb{P}^i\text{-a.e.}). \tag{4.70}$$

Since the above property holds for all such $\hat{\tau}$, it extends to the essential supremum, yielding

$$\rho^i(t_0) \leq \rho^i(t_0) - \varepsilon \; (\mathbb{P}^i\text{-a.e.}), \tag{4.71}$$

or equivalently that $\varepsilon \leq 0$ ($\mathbb{P}^i$-a.e.). In conjunction with the first observation, that $\varepsilon \geqslant 0$ ($\mathbb{P}^i$-a.e.), we have $\varepsilon = 0$ ($\mathbb{P}^i$-a.e.), i.e.,

$$\operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i \left[X^i_{\sigma^i \wedge \hat{\tau}} - X^i_{t_0} - \rho^i(t_0) \mathbb{1}_{\{t_0 < \sigma^i \leq \hat{\tau}\}} \big| \mathscr{F}^i(t_0) \right] = 0 \; (\mathbb{P}^i\text{-a.e.}). \tag{4.72}$$

Define $Z^i_t = X^i_{\sigma^i \wedge t} - X^i_{t_0} - \rho^i(t_0) \mathbb{1}_{\{t_0 < \sigma^i \leq t\}}$. Note that the integrability condition of Lemma 9 is satisfied due to Eq. (4.1). For $t \geqslant \sigma^i$, Z^i_t is constant, hence $\tau^* \leq \sigma^i < \infty$ almost surely. Hence we may apply Lemma 9 here to yield a stopping time $\tau^* \in \hat{\mathbb{F}}^i(t_0)$ such that

$$\mathbb{E}^i \left[X^i_{\sigma^i \wedge \tau^*} - X^i_{t_0} - \rho^i(t_0) \mathbb{1}_{\{t_0 < \sigma^i \leq \tau^*\}} \big| \mathscr{F}^i(t_0) \right] = 0 \; (\mathbb{P}^i\text{-a.e.}), \tag{4.73}$$

or

$$\rho^i(t_0) = \frac{\mathbb{E}^i \left[X^i_{\sigma^i \wedge \tau^*} - X^i_{t_0} \big| \mathscr{F}^i(t_0) \right]}{\mathbb{P}^i \left(t_0 < \sigma^i \leq \tau^* \big| \mathscr{F}^i(t_0) \right)} = \rho^i(t_0, \tau^*) \; (\mathbb{P}^i\text{-a.e.}). \tag{4.74}$$

Hence, the solo-payout index $\rho^i(t_0)$ is realized ($\mathbb{P}^i$-a.e.) for some $\mathbb{F}^i$-stopping time $\tau^* > t_0$.

Proof. [of Proposition 6.] For $k > 0$, let $\tau_k^i < \sigma^i$, and therefore $\tau_{k-1}^i < \sigma^i$. Defining

$$Z_t^i = X_{\sigma^i \wedge t}^i - X_{\tau_{k-1}^i}^i - \rho^i(\tau_{k-1}^i)\mathbb{1}_{\{\tau_{k-1}^i < \sigma^i \leq t\}},$$

note that for $t > \tau_k^i$: $Z_t^i - Z_{\tau_k^i}^i = X_{\sigma^i \wedge t}^i - X_{\tau_k^i}^i - \rho^i(\tau_{k-1}^i)\mathbb{1}_{\{\tau_k^i < \sigma^i \leq t\}}$.

It follows from the proof of Proposition 5 that the solo-payout index from time τ_{k-1}^i is realized by a τ_k^i such that

$$\operatorname*{ess\,sup}_{\tau' \in \hat{\mathbb{F}}^i(\tau_k^i)} \mathbb{E}^i\left[Z_{\tau'}^i \middle| \mathscr{F}^i(\tau_k^i)\right] \leq Z_{\tau_k^i}^i \text{ } (\mathbb{P}^i\text{-a.e.}), \tag{4.75}$$

or

$$\operatorname*{ess\,sup}_{\tau' \in \hat{\mathbb{F}}^i(\tau_k^i)} \mathbb{E}^i\left[X_{\sigma^i \wedge \tau'}^i - X_{\tau_k^i}^i - \rho^i(\tau_{k-1}^i)\mathbb{1}_{\{\tau_k^i < \sigma^i \leq \tau'\}} \middle| \mathscr{F}^i(\tau_k^i)\right] \leq 0 \text{ } (\mathbb{P}^i\text{-a.e.}). \tag{4.76}$$

From the above, for any $\tau' \in \hat{\mathbb{F}}^i(\tau_k^i)$, we have

$$\frac{\mathbb{E}^i\left[X_{\sigma^i \wedge \tau'}^i - X_{\tau_k^i}^i \middle| \mathscr{F}^i(\tau_k^i)\right]}{\mathbb{P}^i\left(\tau_k^i < \sigma^i \leq \tau' \middle| \mathscr{F}^i(\tau_k^i)\right)} \leq \rho^i(\tau_{k-1}^i) \text{ } (\mathbb{P}^i\text{-a.e.}). \tag{4.77}$$

Taking the essential supremum over such τ' establishes that $\rho^i(\tau_k^i) \leq \rho^i(\tau_{k-1}^i)$, ($\mathbb{P}^i$-a.e.).

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] D. Agarwal, B.-C. Chen, P. Elango, and R. Ramakrishnan. Content recommendation on web portals. *Communications of the ACM*, 56(6):92–101, 2013.

[3] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. *arXiv preprint arXiv:1209.3353*, 2012.

[4] A. Arlotto, S. E. Chick, and N. Gans. Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, 60(1):110–129, 2013.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235 – 256, 2002.

[6] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[7] S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. *arXiv preprint arXiv:1302.1611*, 2013.

[8] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. arXiv preprint arXiv:1202.4473, 2012.

[9] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122 – 142, 1996.

[10] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222 – 255, 1997.

[11] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback - Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516 – 1541, 2013.

[12] W. Cowan, J. Honda, and M. N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research, to appear; preprint arXiv:1504.05823*, 2015.

[13] W. Cowan and M. N. Katehakis. Asymptotically optimal sequential experimentation under generalized ranking. *arXiv preprint arXiv:1510.02041*, 2015.

[14] W. Cowan and M. N. Katehakis. An asymptotically optimal UCB policy for uniform bandits of unknown support. *arXiv preprint arXiv:1505.01918*, 2015.

[15] W. Cowan and M. N. Katehakis. Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences*, 29(01):51 – 76, 2015.

[16] A. P. Dawid and M. Stone. The functional-model basis of fiducial inference. *The Annals of Statistics*, pages 1054–1067, 1982.

[17] R. Debouk, S. Lafortune, and D. Teneketzis. On an optimization problem in sensor selection*. *Discrete Event Dynamic Systems*, 12(4):417 – 445, 2002.

[18] i. Dumitriu, P. Tetali, and P. Winkler. On playing golf with two balls. *SIAM Journal on Discrete Mathematics*, 16:604–615, 2003.

[19] B. Efron. RA Fisher in the 21st century. *Statistical Science*, pages 95–114, 1998.

[20] S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning based on Kullback Leibler divergence. In *48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[21] R. G. Frank and R. J. Zeckhauser. Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices. *Journal of Health Economics*, 26(6):1101–1127, 2007.

[22] D. M. N. P. I. Frazier, D. Negoescu, and W. B. Powell. Optimal learning for drug discovery in ewing's sarcoma. 2009.

[23] E. Frostig and G. Weiss. Four proofs of Gittins' multiarmed bandit theorem. *Cyrus Derman Memorial Volume II: Optimization under Uncertainty: Costs, Risks and Revenues* (M.N. Katehakis, S.M. Ross, and J. Yang, eds.), Annals of Operations Research, Springer, 2014.

[24] S. Ghamami, S. M. Ross, et al. Improving the Asmussen–Kroese-type simulation estimators. *Journal of Applied Probability*, 49(4):1188–1193, 2012.

[25] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, (J. Gani, ed.), 241–66, 1974.

[26] J. C. Gittins, K. Glazebrook, and R. R. Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, West Sussex, U.K., 2011.

[27] K. Glazebrook, D. Hodge, and C. Kirkbride. General notions of indexability for queueing control and asset management. *The Annals of Applied Probability*, 21(3):876 – 907, 2011.

[28] K. Glazebrook, C. Kirkbride, H. Mitchell, D. Gaver, and P. Jacobs. Index policies for shooting problems. *Operations research*, 55(4):769 – 781, 2007.

[29] K. D. Glazebrook. On randomized dynamic allocation indices for the sequential design of experiments. *J. R. Statist. Soc. B*, 42:342 – 46, 1980.

[30] K. D. Glazebrook. Optimal strategies for families of alternative bandit processes. *IEEE T. Automat. Contr.*, 28(8):858 – 61, 1983.

[31] K. D. Glazebrook. Methods for evaluating strategies for families of alternative bandit processes. *Journal of Organizational Behaviour and Statistics*, 2(1):1 – 18, 1985.

[32] K. D. Glazebrook and D. M. Jones. Some best possible results for a discounted one - armed bandit. *Metrika*, 30:109 – 15, 1983.

[33] P. Guan, M. Raginsky, and R. M. Willett. Online Markov decision processes with Kullback–Leibler control cost. *Automatic Control, IEEE Transactions on*, 59(6):1423–1438, 2014.

[34] J. Hannig. On generalized fiducial inference. *Statistica Sinica*, pages 491–544, 2009.

[35] J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361 – 391, 2011.

[36] J. Honda and A. Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. *arXiv preprint arXiv:1311.1894*, 2013.

[37] K. Johnson, D. Simchi-Levi, and H. Wang. Online network revenue management using Thompson sampling. *Available at SSRN.*, 2015.

[38] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *3rd international conference on Signals, Circuits and Systems (SCS)*, 2009.

[39] Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1238–1246, 2013.

[40] M. N. Katehakis and C. Derman. Computing optimal sequential allocation rules. In *Clinical Trials*, volume 8 of *Lecture Note Series: Adoptive Statistical Procedures and Related Topics*, pages 29 – 39. Institute of Math. Stats., 1986.

[41] M. N. Katehakis and H. Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

[42] M. N. Katehakis and A. F. Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Math. Oper. Res.*, 12:262 – 68, 1987.

[43] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199 – 13. Springer, 2012.

[44] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 2, 1985.

[45] T. Lattimore and R. Munos. Bounded Regret for Finite-Armed Structured Bandits *Advances in Neural Information Processing Systems*, pages 550–558. 2014.

[46] B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

[47] D. V. Lindley. Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1958.

[48] K. Liu, Q. Zhao, and B. Krishnamachari. Dynamic multichannel access with imperfect channel state detection. *Signal Processing, IEEE Transactions on*, 58(5):2795 – 808, 2010.

[49] J. Mary, R. Gaudel, and P. Preux. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data*, pages 325–336. Springer, 2015.

[50] B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13(1):2069–2106, 2012.

[51] A. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, 1996.

[52] D. M. Negoescu. Bayesian learning models for adaptive treatment decisions in the presence of noisy feedback and treatment-dependent rate of relapses: an application to multiple sclerosis. In *The 35th Annual Meeting of the Society for Medical Decision Making*. Smdm, 2013.

[53] D. M. Negoescu. *Managing Uncertainty in Sequential Medical Decision Making*. PhD thesis, Stanford University, 2014.

[54] D. M. Negoescu, K. Bimpikis, M. L. Brandeau, D. A. Iancu, et al. Dynamic learning of patient response types: An application to treating chronic diseases. Technical report, 2014.

[55] J. Nino-Mora. Stochastic scheduling stochastic scheduling. In *Encyclopedia of Optimization*, pages 3818–3824. Springer, 2008.

[56] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 558:62–76, 2014.

[57] I. Osband and B. Van Roy. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604 – 612, 2014.

[58] Y. Ouyang and D. Teneketzis. On the optimality of myopic sensing in multi-state channels. *arXiv preprint arXiv:1305.6993*, 2013.

[59] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[60] H. Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Monthly*, 58:527–536, 1952.

[61] S. M. Ross. *Introductory statistics*. Academic Press, 2005.

[62] S. M. Ross. *Simulation, Fifth Edition*. Academic Press, 2013.

[63] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[64] A. Singla and A. Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1167 – 1178. International World Wide Web Conferences Steering Committee, 2013.

[65] J. L. Snell. ”Applications of martingale system theorems.” *Transactions of the American Mathematical Society*, 73:293–312, 1952.

[66] I. M. Sonin. Optimal stopping of Markov chains and three abstract optimization problems. *Stochastics An International Journal of Probability and Stochastic Processes*, 83:405–414, 2011.

[67] C. Tekin and M. Liu. Approximately optimal adaptive learning in opportunistic spectrum access. In *INFOCOM, 2012 Proceedings IEEE*, pages 1548 – 1556. IEEE, 2012.

[68] C. Tekin and M. Liu. Online learning of rested and restless bandits. *Information Theory, IEEE Transactions on*, 58(8):5588–5611, 2012.

[69] A. Tewari and P. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems*, pages 1505 – 1512, 2008.

[70] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285 – 94, 1933.

[71] L. Tran-Thanh, L. C. Stavrogiannis, V. Naroditskiy, V. Robu, N. R. Jennings, and P. Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions, 2014.

[72] J. N. Tsitsiliks. A short proof of the Gittins index theorem. *The Annals of Applied Probability*, 194–199, 1994.

[73] W. Wang, S. Lafortune, A. Girard, and F. Lin. Optimal sensor activation for diagnosing discrete event systems. *Automatica*, 46(7):1165 – 1175, 2010.

[74] Z. Wang, S. Deng, and Y. Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318 – 331, 2014.

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

[76] J. White. *Bandit algorithms for website optimization.* " O'Reilly Media, Inc.", 2012.

[77] P. Whittle. Multi-armed bandits and the Gittins index. *J. Royal Statistical Society Series B*, 42:143–149, 1980.