High Frequency Sampling in Bandit Problems with Dynamically Changing Reward Distributions

Evan Sadler

April 30, 2009
Contents

1 Introduction 3

2 Motivating Applications 6
   2.1 Market Learning ................................. 6
   2.2 Treatment Allocation ............................ 9
   2.3 Marketing ....................................... 9

3 Preliminary Work 10

4 The Dynamic Model 15
   4.1 First Approach .................................. 17
   4.2 Generalized Upper Confidence Bound Policies .......... 23
   4.3 Optimal Estimation ................................. 27
   4.4 Refined Dynamic UCB Allocation Policies ............... 31
   4.5 Non-Uniform Arrival Times .......................... 33

5 Simulations 35
   5.1 Comparison Between Policies ........................ 37
   5.2 Sensitivity to Noise ................................ 37
   5.3 Multiple Arms .................................... 38

Appendices 38

A Comparisons 38

B Sensitivity to Noise 55

C Multiple Arms 58
1 Introduction

A bandit problem is a sequential allocation problem in which an agent faces a finite number of actions to choose between, and each is associated with a particular reward. Choosing an action allows the agent to realize a reward, typically with a certain degree of noise preventing an exact observation. A defining characteristic of a bandit problem situation is that rewards from multiple actions cannot be realized or observed simultaneously. This aspect leads to the dilemma at the heart of any bandit problem: should the empirically best action be chosen (exploitation) or should more information be gathered on the rewards of other actions in the hope that previous observations were simply unlucky (exploration)? The exploitation versus exploration dilemma is a fundamental problem in several areas—statistics, operations research, and economics for instance—and consequently the bandit problem model has far reaching applications.

The basic (static) model of a multi-armed bandit problem can be described as follows: There are $M$ gambling machines each associated with a random variable $X_i$, $1 \leq i \leq M$, from which rewards are drawn. Successive plays of the same machine yield a stream of independent and identically distributed rewards. Independence holds across machines as well, but distributions will usually not be identical. Reward distributions are unknown a priori, but information can be gathered by playing a machine and realizing a reward.

A solution to a multi-armed bandit problem is an allocation policy: an algorithm that decides what machine to play next given all previous plays and realized rewards. We can measure the quality of any particular allocation policy over a finite number, $n$, of plays as the expected loss due to the fact that we are not always playing the optimal machine. If $\mu_i$ denotes the expected value of the random variable $X_i$, and $\mu^*$ the expected reward from the best machine, then this expected loss after $n$ plays—often referred to as the regret—is given by

$$\mu^* n - \sum_{i=1}^{M} \mu_i E[\#_i(n)]$$

Here, $\#_i(n)$ denotes the number of times machine $i$ is played during the first $n$ plays.

Robbins (1952) was the first to study sequential allocation problems using a model arising from the problem of optimal population sampling. In his brief treatment, Robbins introduced the idea of measuring the performance of an allocation strategy via the expected deviation from optimal allocation under full information and was able to show that under very general circumstances, it
is possible to devise a strategy for which the asymptotic loss per action due to incomplete information approaches zero as the number of actions taken goes to infinity.

Since that time, bandit problems have been considered in a variety of contexts and applications, including the following examples:

**Market Learning**
A firm facing unknown demand for its products can sequentially adjust prices to gain information about the market it faces. Once a price is set over a period, the firm makes a noisy observation of the market demand. The goal is to find the profit maximizing price as quickly as possible (see Rothschild, 1974).

**Clinical Trials**
Closely related to the population sampling problem which motivated Robbins' original paper, patients can sequentially receive one of a set of treatments. A noisy measurement of the efficacy is observed, and the goal is to give the best treatment to as many patients as possible (see Bather, 1985).

**Marketing**
Marketers sequentially send advertising messages to consumers (through banners on a web page for instance), but the real efficacy of each message is unknown. Various noisy indicators of efficacy can be observed, however, and the messages sent can be adjusted to maximize the generation of sales (see Gooley and Lattin, 2000).

It took more than 30 years to develop policies that provably achieved optimal regret growth for the static problem. In a seminal paper, Lai and Robbins (1985) demonstrated policies that achieve total regret growing with the logarithm of the number of plays, and proved that this is the best possible asymptotic result in the static problem. Lai and Robbins' allocation policies were based on the idea of computing upper confidence bounds for each machine and choosing the machine with the highest bound. Thus, these policies are choosing the machine with the highest potential reward in a specific sense. Lai and Robbins made use of parametric assumptions on the reward distributions in order to obtain the optimal regret bounds. More recently these assumptions have been relaxed; Auer et al. (2002) assume that the only knowledge available a priori about the reward distributions is that they are bounded almost surely.

The early models, as well as the vast majority of extant literature on bandit problems, deal with sequential allocation in a static context. That is, every time a particular action is taken the distribution of rewards remains the same. Clearly, strategies derived from these models may fail to produce good results in a dynamic environment. The examples mentioned above may not closely match the static model, and indeed, assuming independent and identically distributed rewards from any single action is extremely restrictive when considering real ap-
lications. Relaxing this assumption has the potential to unlock a wealth of new and useful models, but the literature on dynamic bandit models is surprisingly sparse.

In recent years, there has been a renewed interest in sequential allocation—particularly as it relates to machine learning—and a number of new environments have started to be examined. Extensions of the basic model include considering a version of the bandit problem in which each machine is associated with a finite-state Markovian reward process (see Gittins, 1989; Berry and Fristedt, 1985), modeling continuously changing machine rewards as standard Weiner processes (see Slivkins and Upfal, 2008), the introduction of side information or covariates (see Yang and Zhu, 2002), and others (see Garivier and Moulines, 2008). We will examine a new model in which an agent acts in a dynamically changing environment over a finite time horizon, with performance ultimately depending on the frequency with which the agent can sample from the machines. Playing machine $i$ will draw from an unknown, but bounded, distribution with unknown expectation given by the function $f_i(t)$.

An important challenge arising in the transition from a static environment to a dynamic one is the choice of appropriate stability assumptions. Such conditions are needed to ensure that past observations still contain enough information to guide future decisions. In the worst case scenario—a chaotically changing environment—we cannot hope to do better than what is achieved by randomly selecting actions. Our model will make use of three basic assumptions. The first two—Hölder continuity of $f_i$ for each $i$ and bounded reward distributions over the entire time interval—control the variation in machine rewards over time allowing the derivation of estimators for each $f_i$ and confidence bounds on those estimators. The third assumption—the margin condition—controls the proximity of the $f_i$ to each other over the time interval. Take $\delta_0$, $C$, and $\alpha$ to be three positive constants. The bandit problem satisfies the margin condition if for any two machines, $i$ and $j$, we have

$$\lambda\{t : 0 < |f_i(t) - f_j(t)| \leq \delta\} \leq C\delta^\alpha$$

where $\lambda$ denotes Lebesgue measure. These relatively weak assumptions will be sufficient for us to adapt the methodology of upper confidence bound policies to our purpose by borrowing ideas from nonparametric estimation.

Our discourse will be structured in the following manner. Chapter 2 will provide motivating examples of applications, suggesting the breadth of uses for dynamic bandit models. Chapter 3 will provide an overview of the basic multi-armed bandit problem in a static context along with the tools needed to construct upper confidence bound policies. Chapter 4 will introduce our dynamic model and derive relevant results with respect to proposed allocation policies. Chapter 5 will provide simulations and will assess the performance of
the allocation policies in practice.

2 Motivating Applications

This chapter describes several bandit problem applications in which dynamic considerations and the associated allocation policies could broaden the power of the bandit problem model. We consider in sequence each of the examples mentioned in the introduction. In the case of market learning, the suitability of our model to dynamically changing environments with high frequency decision making makes the example of market making in financial securities especially interesting. Allocation of treatments, as for infectious diseases, also lends itself to the dynamic model. We can consider a patient population afflicted by several different, difficult to distinguish, strains of the same pathogen which respond differently to our treatment options. If the proportions of the strains change over time, or if we allow ourselves to consider the development of resistance to particular treatments, allocation policies based on the static model may prove suboptimal. Problems faced by marketers utilizing interactive media—such as web pages and banners—also clearly motivate treating sequential allocation in a dynamic context.

The model we propose will provide a framework through which dynamic features of all of these problems can be considered. The allocation policies we introduce are applicable to each of these situations and have the potential to achieve substantially better results than policies based on a static model.

2.1 Market Learning

Rothschild (1974) introduced an early bandit model of market learning in which a store is attempting to price an item in its inventory. To simplify the analysis, it is assumed that potential customers arrive at a constant rate, regardless of the price, and then decide whether or not to buy. While this is generally not a sound assumption to make—as noted in Rothschild’s paper—there are some markets in which it is not unreasonable. Examples include markets in which repeat business cannot be expected in the near term, stores which sell multiple items so that the pricing of one has little influence on a customer’s decision to visit the store, and markets in which customers do not know the price of an item before entering the store.

A strategy for the store is then determined by selecting a price from a set of prices \( \{ p_i \} \), \( i = 1, 2, \ldots, k \). A potential customer then purchases the item with probability \( \Pi_i \) and depending on the customer’s decision, the store observes a profit of either 0 or \( q_i = p_i - c \) where \( c \) is the item’s cost. In the presence of full information on market demand, the store owner would simply choose \( p_i \) to
maximize $\Pi_q$. In the absence of this information, the store owner must consider the prices chosen not only as they affect the immediate profit of the store, but also as they reveal information about the probabilities $\Pi_i$: every time a price $p_i$ is offered to a potential customer, the store owner makes a noisy observation of the demand. If demand were unchanging over time, an allocation policy tailored to a static bandit model would be optimal. The prices are the arms of the bandit, and there are well understood policies for choosing which arm to pull next that would allow the store owner to lose out on profits proportional only to the logarithm of the number of potential customers.

The assumption of constant demand, however, is unlikely to be a good assumption. For many products, demand will respond to trends in consumer preference, to innovation in the product market, to seasonality, and in some cases even to the time of day (e.g. a last minute run to the grocery store before dinner time). The store owner’s problem is more accurately understood using a dynamic model that accounts for these changes. In practice, the store owner is likely to be able to improve upon a static policy by accounting for the fact that many demand fluctuations are regular over time and are easy to predict (seasonality, time of day). However, other demand fluctuations (trends, innovation) are perpetually unpredictable. Ultimately, the ability of any particular seller to adapt to dynamically changing market demand will depend primarily upon two factors: the frequency with which they can make observations of demand, and the persistence of market demand (i.e. the rate at which demand changes over time). Sellers in the financial markets—who face a very dynamic demand environment and can continually update their prices—offer a particularly extreme real world example in which a static bandit model is clearly inadequate, but the high frequency of observations could allow an allocation policy tailored to a dynamic environment to do quite well.

Consider a market maker trading in a particular security over the course of a day. At any particular moment, this market maker can publish a single price at which they will offer to sell the security. The market maker wishes to price their securities so as to maximize profit, but successful implementation assumes a dauntingly high level of knowledge about the market. The profit-maximizing market maker must know perfectly the level of demand for the security at all times throughout the day and the price sensitivity of all actors in the market for the security.

Clearly, expecting such knowledge of any market maker is unreasonable, even more so than expecting a general store owner to know perfectly the demand for each item in inventory. The fast pace of the financial markets severely complicates the market maker’s allocation problem. Developing a useful pricing scheme for the financial market maker requires making use of only that knowledge which is readily available and of that knowledge, only that which can reasonably be synthesized before a decision must be made. For a market maker trading in a high speed environment, knowledge of demand for a security is obtained only
by directly observing the sales made of that security. At each moment in time, the market maker faces a choice of what price to charge and is rewarded with a particular order flow translating directly into a level of profit. Unfortunately for the market maker, this realized order flow is an imperfect measurement of demand, and demand for a security changes over time. The market maker must extract as much information as possible from these observations in order to optimally set prices.

We can consider a bandit model of this situation by making a few simplifying assumptions on the problem faced by the market maker. First, we assume that the market maker is a price taker in the supply market for securities. Under this assumption, the market maker’s choice of price is equivalent to a choice of the bid-ask spread. The strategies available to the market maker are thereby confined to a relatively small set, and the market maker’s choice is dissociated from the absolute price level in the market, which can vary quite drastically. Second, we assume that the pricing strategies of other market makers are exogenous. In other words, there is no feedback mechanism facilitating other market makers to change their prices in response to our pricing scheme. If our market maker is not so large as to exert significant market power, this should be a safe assumption. We will initially follow Rothschild’s model in assuming that potential customers arrive at a constant rate. Later we will generalize our bandit problem model to allow for a variable arrival rate. Unlike in Rothschild’s model, we will allow that a single customer may purchase several shares of our security upon visiting our market maker, but we suppose that the number of shares a customer may purchase is bounded.

We can now fit this problem to a dynamic bandit model. As a customer arrives, they are quoted a price $p_i$ and choose a non-negative number of shares to purchase drawn from a random variable distribution with expectation given by the function $S_i(t)$, representing demand for the security at time $t$. We will assume that demand $S_i(t)$ is Hölder continuous over the trading day and satisfies the margin condition. Let $q_i$ denote the profit per share for the market maker using pricing scheme $p_i$; the problem is now to maximize $S_i(t)q_i$. Our market maker may not be able to perfectly price the security in question, but the observed results of previous pricing decisions can aide the determination of future prices to offer. Finding a good pricing scheme depends upon understanding the behavior of allocation policies under this kind of dynamic environment.

A few difficulties may arise in practice for a market maker attempting to implement a pricing strategy based upon this model. The first, and most obvious, stems from our assumption that $S_i(t)$ is Hölder continuous. Anyone who studies the markets realizes that demand is subject to occasional shocks, which would manifest themselves as jumps in the reward expectation function $S_i(t)$. A policy based upon our model would be able to retain its regret growth if we could recognize these shocks to the markets and restart. For some shocks, such as responses to major world events or disasters, this is not entirely unreasonable.
Unobservable shocks, however, could pose major problems. A second difficulty also stems from shocks to the market, but its effect on our policies is less direct. To the extent that our allocation policies depend on the bounds we can place on the number of shares a customer will purchase, demand shocks, and especially large swings in market volume, may require tuning of some parameters.

2.2 Treatment Allocation

Suppose we are faced with a sequence of patients, each of whom suffers from a common disease for which we have two viable treatments. To simplify this model, we further suppose that success or failure of the treatment can immediately be assessed. In reality there may be a substantial time delay before this information can be known, and this may cause our model to overstate the advantages of sequential allocation methods, but there is still a clear benefit from using the information obtained through these techniques as it becomes available. We assume that the treatments are not equally effective, and that the efficacy of each treatment over time is not constant. For the purposes of our model, we assume that the factors influencing the changing efficacy of the treatments are exogenous. They could be related to mutating strains of the disease, changing characteristics of the infected population, or a variety of other factors. We want to allocate treatments to these patients so as to minimize the failure rate of treatment.

Likely, not all patients will be willing participants in sequential experimentation on the treatment options, but we can assume that some non-empty subset will volunteer. At any given point in time \( t \), treatment \( i \) has a probability of success given by \( f_i(t) \) for which we will want an accurate estimate \( \hat{f}_i(t) \). Patients in the volunteer group can be given treatment according to a sequential allocation policy, while those outside the volunteer group should be given the treatment with the current highest estimated success rate. The work laid out in the next two chapters will address the problems we face with both of these groups by exploring refined techniques for estimating \( f_i(t) \) as well as allocation policies.

2.3 Marketing

Interactive media has revolutionized marketing and has created new applications for sequential allocation models. Prior to the advent of interactive media, in particular the internet, marketers were forced to send a standardized message to all potential consumers, which could only be updated as often as a particular publication is produced. As noted by Gooley and Lattin (2000), interactive media allows the marketer to identify characteristics of the consumer, customize an ad message in real-time, and collect response data. The marketer’s objective is to maximize response rate by dynamically allocating a set of unique
marketing messages to potential consumers. Our bandit model will address the
effect of changing responses over time on allocation, but the methods could be
generalized to deal with other covariates impacting consumer response.

Suppose a marketer has two or more unique messages that can be delivered
to potential customers who arrive sequentially at a web page. Further, suppose
that the marketer has a reliable method of measuring the response of potential
customers (e.g. clicks on the ad banner, actual purchases, time spent on the web
page) and for each customer the response to ad $i$ is a realization of a random
variable with expectation $f_i(t)$. For the purposes of our dynamic bandit model
we will assume that these expectations are Hölder continuous functions of time
and satisfy the margin condition. The marketer can learn about the efficacy
of the messages by observing the responses of potential customers, and makes
future decisions about what message to display based upon the past sequence
of messages shown and the observed responses. A dynamic allocation policy
based upon this model has the potential to significantly improve the marketer’s
ability to generate favorable responses to an ad campaign.

The examples laid out in this chapter should demonstrate that applications
of dynamic time-dependent bandit models are numerous, and the continued de-
velopment of technology allowing high speed trading, information gathering, and
interactive media only makes these models more relevant. The next chapter will
initiate our work on developing policies for our dynamic model by introducing
upper confidence bound policies through the static problem.

3 Preliminary Work

Consider an $M$-armed bandit as presented in the introduction in which each
arm has an arbitrary, fixed reward distribution taking values on $[0,1]$ almost
surely. Our allocation policy will consist of two parts: first, an initialization
period in which each machine is played once—and second, a continuation period
in which new upper confidence bounds are computed for each machine after
every play, and the machine with the highest bound is played next. For a given
machine $i$, these bounds before the $n + 1$th play will be computed as functions
of the form

$$X_{i, \#_i(n)} + c_{n, \#_i(n)}$$

Here, $X_{i, \#_i(n)}$ is an empirical average reward from machine $i$ over the $\#_i(n)$
observations made during the first $n$ plays, and $c_{n, \#_i(n)}$ is an error term de-
pending on both the total number of plays thus far, and the number of times
machine $i$ has been played.

We make use of our previous notation, letting $\#_i(n)$ denote the number of
times machine i has been played in the first n periods. The desired results for the total regret

$$\mu^*n - \sum_{i=1}^{M} \mu_i E[\#_i(n)]$$

will be derived by placing bounds on $E[\#_i(n)]$. Obtaining optimal growth for the total regret requires the use of an especially tight stochastic inequality due to Hoeffding (1963).

**Lemma 1.** (Hoeffding’s Inequality) Let $X_1, X_2, ..., X_n$ be independent random variables that are bounded almost surely

$$P(X_i \in [a_i, b_i]) = 1$$

$$1 \leq i \leq n; |a_i|, |b_i| < \infty$$

Let $E[X_i] = \mu_i$ and define $\mu = \sum_{i=1}^{n} \mu_i/n, b = \sum_{i=1}^{n} b_i/n$. Then if $0 < \delta < n(b - \mu)$ and $S = \sum_{i=1}^{n} X_i$ we have the following inequalities:

$$P(S - E[S] \geq \delta) \leq e^{-\frac{2\delta^2}{\sum_{i=1}^{n}(a_i - b_i)^2}}$$

$$P(|S - E[S]| \geq \delta) \leq 2e^{-\frac{2\delta^2}{\sum_{i=1}^{n}(a_i - b_i)^2}}$$

We will make a quick note regarding the assumption made on $\delta$. Notice that under the lemma’s assumptions regarding the boundedness of the random variables $X_i$, if $\delta > n(b - \mu)$ the probability on the left is simply zero. In the special case of equality, the lemma remains true by replacing the right hand side with the limit as $\delta$ approaches $n(b - \mu)$ from the left.

**Proof.** (Adapted from Cesa-Bianchi and Lugosi, 2006) We begin with a simple observation: $P(S - E[S] \geq \delta) = E[Y]$, where $Y$ is a random variable that takes the value 0 when $S - E[S] - \delta < 0$ and the value 1 otherwise. Trivially, for any positive value $h$, we have

$$Y \leq e^{h(S - E[S] - \delta)}$$

implying

$$P(S - E[S] \geq \delta) \leq E[e^{h(S - E[S] - \delta)}] = e^{-h\delta} \prod_{i=1}^{n} E[e^{h(X_i - \mu_i)}]$$
since the \( X_i \) are independent.

Another simple fact we will use derives from Jensen’s inequality and the fact that the exponential function is convex. For a random variable \( X \) bounded between \( a \) and \( b \) we have:

\[
E[e^{hX}] \leq \frac{b - E[X]}{b - a} e^{ha} + \frac{E[X] - a}{b - a} e^{hb}
\]

For each index \( i \) let \( p_i = \frac{a_i}{b_i - a_i} \). Then using the above inequality and the fact that \( E[X_i - \mu_i] = 0 \) we can rewrite our bound on \( P(S \geq \delta) \) as

\[
P(S - E[S] \geq \delta) \leq e^{-h\delta} \prod_{i=1}^{n} e^{h(X_i - \mu_i)}
\]

We can write each term of this last product as \( e^{G_i(\nu)} \) where \( \nu_i = h(b_i - a_i) \) and we define

\[
G_i(\nu) = -p_i \nu + \ln (1 - p_i + p_i e^\nu)
\]

We compute

\[
G'_i(\nu) = -p_i + \frac{p_i}{p_i + (1 - p_i)e^{-\nu}}
\]

and

\[
G''_i(\nu) = \frac{p_i(1 - p_i)e^{-\nu}}{(p_i + (1 - p_i)e^{-\nu})^2} \leq \frac{p_i(1 - p_i)e^{-\nu}}{4p_i(1 - p_i)e^{-\nu}} = \frac{1}{4}
\]

by the arithmetic-mean—geometric-mean inequality.

Thus, we have that for each \( i \), \( G_i(0) = G'_i(0) = 0 \) and by Taylor’s theorem

\[
G_i(\nu) = G_i(0) + \nu G'_i(0) + \frac{\nu^2}{2} G''_i(\theta) \leq \frac{\nu^2}{8}
\]

for some \( \theta \in [0, \nu] \). Therefore, we have

\[
P(S - E[S] \geq \delta) \leq e^{-h\delta} \prod_{i=1}^{n} e^{\nu_i^2} \leq e^{-h\delta} \prod_{i=1}^{n} e^{\frac{h^2(b_i - a_i)^2}{8}}
\]
for every positive $h$. Minimizing over $h$ gives $h = \frac{4\delta}{\ln n - a_i \pi^2}$, and the result follows by substituting this value of $h$, completing our proof.

With this result in hand we can proceed with our work on the static bandit problem originally described. The proof for the regret bound on the policy laid out in this section is due to Auer et al. (2002).

### Static UCB Policy

- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ always play that machine which maximizes $X_i, \#_i(n) + c_{n, \#_i(n)}$, where $c_{t, s} = \sqrt{\frac{2 \ln t}{s}}$

#### Theorem 2. The Static UCB Policy achieves expected regret bounded by

\[
\sum_{i \neq \star} \left[ \frac{8 \ln(n)}{(\mu^* - \mu_i)} + (\mu^* - \mu_i) \left( 1 + \frac{\pi^2}{3} \right) \right]
\]

**Proof.** Let $I_t$ denote the index of the machine played at time $t$, and in the following sums, let $\chi(Statements)$ denote the indicator function which evaluates to 1 if each statement in the brackets is true, and 0 otherwise. Thus, given the description of our allocation policy we begin by writing

\[
\#_i(n) = 1 + \sum_{t=M+1}^{n} \chi(I_t = i) \leq l + \sum_{t=M+1}^{n} \chi(I_t = i, \#_i(t-1) \geq l)
\]

Let $\overline{X}_{\#^*(t)}$ denote the empirical average of the best machine, using data only from the first $t$ plays, and similarly let $\#^*(t)$ denote the number of times the best machine has been played during the first $t$ plays. The condition $I_t = i$ can only be met if

\[
\overline{X}_{\#^*(t-1)} + c_{t-1, \#^*(t-1)} \leq \overline{X}_{i, \#_i(t-1)} + c_{t-1, \#_i(t-1)}
\]

so we can rewrite our bound as
\[
\#_i(n) \leq l + \sum_{t=M+1}^{n} \chi(\overline{X}_{#_i}(t-1) + c_{t-1, #_i}(t-1) \leq \overline{X}_{i, #_i}(t-1) + c_{t-1, #_i}(t-1), \#_i(t-1) \geq l)
\]

Likewise, these two conditions are simultaneously met only if
\[
\min_{0 < s < t} X^*_s + c_{t-1, s} \leq \max_{t \leq s < t} X_{i,s} + c_{t-1, s}
\]

Here, \( \overline{X}_{i,s} \) is taken to be an average over the first \( s \) observations of the random variable \( X_i \). We can now derive the following crude bound on \( \#_i(n) \):

\[
\#_i(n) \leq l + \sum_{t=M+1}^{n} \chi(\min_{0 < s < t} X^*_s + c_{t-1, s} \leq \max_{t \leq s < t} X_{i,s} + c_{t-1, s})
\]

\[
\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=t}^{t-1} \chi(X^*_s + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s_i})
\]

One of three things must be true if \( X^*_s + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s_i} \):

1. \( X^*_s \leq \mu^* - c_{t,s_i} \),
2. \( \overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \), or
3. \( \mu^* < \mu_i + 2c_{t,s_i} \).

The function, \( c \), was chosen to bound the probability of the first two in such a way that the corresponding triple sum in our inequality converges to a constant. We will choose \( l \) so that the probability of the third possibility is zero. Using lemma 1 we note that our choice of \( c_{t,s} = \sqrt{2 \ln(t}/s_i \) yields

\[
P(\overline{X}^*_s \leq \mu^* - c_{t,s}) \leq t^{-4}
\]

\[
P(\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i}) \leq t^{-4}
\]

Choosing \( l \geq 8 \ln(n)/\left(\mu^* - \mu_i\right)^2 \) gives

\[
\mu^* - \mu_i - 2\sqrt{2 \ln(t}/s_i \geq \mu^* - \mu_i - (\mu^* - \mu_i)\sqrt{\ln(t}/\ln(n)}
\]

\[
\geq \mu^* - \mu_i - (\mu^* - \mu_i)
\]

\[
\geq 0
\]
and combining these bounds gives
\[
E[\#(n)] \leq \frac{8 \ln(n)}{(\mu^* - \mu_i)^2} + 1 + \frac{\pi^2}{3}
\]
so total expected regret from this allocation policy is bounded as
\[
\sum_{i \neq \ast} \left[ \frac{8 \ln(n)}{(\mu^* - \mu_i)} + (\mu^* - \mu_i) \left( 1 + \frac{\pi^2}{3} \right) \right]
\]

The optimal logarithmic regret growth is thus obtained by the allocation policy which plays each machine once to initialize, continues by computing an index for each machine \( i \) at time \( t \) of \( X_i + \sqrt{2 \ln(t) / \#(t)} \), and always plays the machine with highest index. The desired bound is easy to achieve in this case because the rewards from each machine are i.i.d. This allows us to sum over arbitrary averages without regard to the ordering of the observations. We will make use of this policy in our treatment of a dynamic problem in the next chapter, and will later attempt to generalize the upper confidence bound approach and modify the policy in an effort to improve our performance in dynamic problems.

4 The Dynamic Model

We consider a bandit problem over the finite time period \([0, 1]\) and an agent choosing an action after each \( 1/N \) units of time. There are \( M \) machines available to the agent during each of these time periods, and at time \( t \) machine \( i \) yields a reward realized from a random variable supported on \([0, 1]\) almost surely with expectation \( f_i(t) \), \( i = 1, 2, ..., M \). For each \( i \), \( f_i(t) \) is Hölder continuous:
\[
|f_i(t_1) - f_i(t_2)| \leq L|t_1 - t_2|^\beta
\]
\[
0 < \beta \leq 1
\]
for some fixed \( \beta \) and some constant \( L \). Finally, we have the margin condition. Take \( \delta_0 \), \( C \), and \( \alpha \) to be three positive constants. The bandit problem satisfies the margin condition if for any two machines, \( i \) and \( j \), we have
\[
\lambda \{ t : 0 < |f_i(t) - f_j(t)| \leq \delta \} \leq C\delta^\alpha
\]
for any \( t \in [0, 1] \) and any positive \( \delta \) with \( \delta < \delta_0 \), where \( \lambda \) denotes Lebesgue measure.
We should note that our smoothness condition and the margin condition are working against each other as they control the functions \( f_i \). Given two machines with expected rewards \( f_i \) and \( f_j \) which intersect, Hölder continuity ensures that their difference does not grow away from zero too quickly, while the margin condition does the opposite. The conflict between the parameters \( \beta \) and \( \alpha \) is characterized in the following proposition:

**Proposition 3.** If there does not exist an optimal allocation policy which is constant (i.e. the policy always plays the same machine), then \( \alpha \beta \leq 1 \).

We make a quick remark about the assumption in the proposition statement. There are two different ways to have a constant allocation policy which is optimal. In one case, there is simply a single machine whose expected reward is always higher than that of all other machines. The other possibility is that there are multiple machines whose expected rewards are always equal and greater than the expectations of all other machines. For there to be no optimal allocation policy which is constant, there must be two machines whose reward expectation functions intersect and move apart.

**Proof.** Consider two such intersecting functions, \( f_i \) and \( f_j \), which satisfy both the Hölder continuity condition and the margin condition, and let \( \lambda \) denote Lebesgue measure. Since \( f_i \) and \( f_j \) are Hölder continuous with parameter \( \beta \) and constant \( L \), their difference is Hölder continuous with parameter \( \beta \) and constant \( 2L \). We have assumed that \( f_i \) and \( f_j \) are not everywhere equal, so without loss of generality we now assume that there exists a time, \( t^* \), such that \( f_i(t^*) = f_j(t^*) \) and \( f_i(t) \neq f_j(t) \) for \( t \) in an interval \( [t^*, t^* + \delta^*] \). Taking any \( \delta < \delta^* \), by Hölder continuity we have

\[
|f_i(t) - f_j(t)| < 2L\delta^\beta
\]

for \( t \) in the interval \( [t, t + \delta] \). This fact implies that for any sufficiently small \( \delta \) it must be the case that

\[
\delta \leq \lambda\{t : 0 < |f_i(t) - f_j(t)| < 2L\delta^\beta\}
\]

If \( \delta \) is selected small enough so that \( 2L\delta^\beta \leq \delta_0 \) we then have by the margin condition

\[
\lambda\{t : 0 < |f_i(t) - f_j(t)| < 2L\delta^\beta\} < C\lambda \delta^{3\alpha}
\]

If \( \alpha \beta > 1 \), we can choose \( \delta \) small enough so that \( \delta^{\alpha \beta} < \delta/(C\lambda) \), contradicting the inequalities just derived. Thus, \( \alpha \beta \leq 1 \).
In this chapter we will devise upper confidence bound policies for the problem just described. Just as in the static context, each of our policies will involve assigning an index to each machine that is the sum of an estimator for its expectation, and a term representative of the uncertainty in that estimation. Unlike the static context, a good estimator cannot simply be an average over all past rewards. To derive suitable confidence bounds, the estimator will need to account for the uncertainty resulting from the time which has elapsed since our observations of a machine’s reward. Our development of policies for this dynamic bandit problem will begin by taking a simple adaptation of the static upper confidence bound policy and deriving regret bounds attainable with this adaptation. Potential means to improve the basic policy will then be considered. In particular, we will address the problem of estimating a Hölder continuous function based upon high frequency, noisy observations—an interesting problem in its own right. Finally, realizing that relaxing the assumption that the agent faces new choices at constant time intervals is desirable—particularly in applications when the agent’s choice corresponds to the arrival of a potential customer—we will further generalize our model to account for high frequency observations with a non-uniform arrival process.

4.1 First Approach

In our first pass at the dynamic bandit problem outlined in the introduction to this chapter, we will consider the situation as a sequence of $K$ static bandit problems to be approached separately using exactly the same strategy outlined in chapter 3. The bin size, $1/K$, will be chosen as a function of the playing frequency $N$ and the smoothness parameter $\beta$ to balance the benefit of variance reduction due to more observations per bin against the benefit of reduced bias resulting from smaller bin size.

**Dynamic UCB Policy 1**

- Choose the number of bins, $K$
- Begin by implementing the Static UCB Policy
- At the beginning of each time interval $[j/K, (j + 1)/K]$, $j = 1, 2, 3, ..., K - 1$, reset by discarding all historical information and reinitialize the Static UCB Policy as though facing the start of a new bandit problem

Proving a regret bound for this policy will rely on the ability of the Static UCB Policy to retain its performance characteristics in a dynamic setting.
when the reward expectation functions are sufficiently separated. The following lemma assures us of this.

**Lemma 4.** Suppose \( f_1(t) \), and \( f_2(t) \) are any two functions taking values in the interval \([0, 1]\) such that

\[
\inf_t f_1(t) = \mu_1 > \mu_2 = \sup_t f_2(t)
\]

Consider a bandit problem with two machines whose reward distributions are supported on the interval \([0, 1]\) and have expectations given by \( f_1(t) \) and \( f_2(t) \). If we implement the Static UCB Policy on this bandit, the expected number of mistakes we will commit after \( n \) plays is bounded as

\[
\mathbb{E}[\#_2(n)] \leq \frac{8 \ln(n)}{(\mu_1 - \mu_2)^2} + 1 + \frac{\pi^2}{3}
\]

**Proof.** The proof of the lemma largely follows the proof of the regret bound for the static policy. Taking again \( X_{i,s} \) to denote the average observed reward from machine \( i \) over the first \( s \) observations, we can bound the expected number of mistakes using the same triple sum as before

\[
\mathbb{E}[\#_2(n)] \leq l + \sum_{t=1}^{l} \sum_{s=1}^{t-1} \sum_{s_2=l}^{t-1} \chi(X_{1,s} + c_{t,s} \leq X_{2,s_2} + c_{t,s_2})
\]

and for \( X_{1,s} + c_{t,s} \leq X_{2,s_1} + c_{t,s_2} \) to hold, at least one of the same three relationships must be true:

1. \( X_{1,s} \leq \mu_1 - c_{t,s} \),
2. \( X_{2,s_2} \geq \mu_2 + c_{t,s_2} \), or
3. \( \mu_1 < \mu_i + 2c_{t,s_2} \).

As before, choosing \( l \geq 8 \ln(n)/(\mu_1 - \mu_2)^2 \) ensures that the third possibility cannot happen, but bounding the probabilities of the first two requires a little more care in this instance. Hoeffding’s inequality tells us that

\[
P(X_{1,s} \leq \mathbb{E}[X_{1,s}] - c_{t,s}) \leq t^{-4}
\]
\[
P(X_{2,s_2} \geq \mathbb{E}[X_{2,s_2}] + c_{t,s_2}) \leq t^{-4}
\]

Unfortunately, we do not know what the expected values of these averages are because they are dependent on the order and timing of our samples. However,
the bounds we assumed on \( f_1 \) and \( f_2 \) guarantee that regardless of sample order or timing we have

\[
E[\bar{X}_{1,s}] \geq \mu_1
\]

\[
E[\bar{X}_{2,s}] \leq \mu_2
\]

This guarantees that

\[
P(\bar{X}_{1,s} \leq \mu_1 - c_{t,s}) \leq t^{-4}
\]

\[
P(\bar{X}_{2,s} \geq \mu_2 + c_{t,s}) \leq t^{-4}
\]

and the rest of the proof is the same as before.

Using this result, we can achieve a regret bound for Dynamic UCB Policy 1 by independently considering the bins in which expected rewards are well separated and the bins in which expected rewards are close. To simplify our work, we will initially assume that we have only two machines to choose from. The following result describes the regret attainable by implementing Dynamic UCB Policy 1.

**Theorem 5.** Consider a high frequency dynamic two-armed bandit whose reward expectation functions satisfy Hölder continuity and the margin condition. Implementing Dynamic UCB Policy 1 with \( K = C_1 (N/\ln N)^{1/(2\beta+1)} \) will yield expected regret bounded by

\[
C_2 N \left( \frac{N}{\ln N} \right)^{-\beta(1+\min(\alpha,1))^{2\beta+1}}
\]

where \( C_1 \) and \( C_2 \) are positive constants.

The techniques used in the following proof are adapted from Rigollet and Zeevi (2009).

Proof. We index the successive bins as \( B_i, i = 1, 2, ..., K \), and partition them into two classes—those that are well-behaved relative to our upper confidence bound policy, and those that are not. The well-behaved bins will be characterized by the set

\[
\mathcal{I} = \{ i : \forall t \in B_i, |f_1(t) - f_2(t)| > cK^{-\beta} \}
\]
where $c$ is a positive constant chosen to be larger than $3L$, where $L$ is the Hölder continuity constant. Since each bin has length $1/K$, Hölder continuity assures us that we can find another constant $c'$ such that $|f_1(t) - f_2(t)| \leq c'K^{-\beta}$ for any $t \in B_i$ where $i \notin I$. Since $f_1$ and $f_2$ are continuous on a compact set, they are equal to each other on at most a finite number of intervals (including isolated points). Thus, for all values of $N$, there is a constant bound, $b$, on the number of bins $B_i$, $i \notin I$, that contain both points with $f_1 = f_2$ and points with $f_1 \neq f_2$. Let $J$ denote the set of bins which are not in $I$ and for which $f_1$ is not uniformly equal to $f_2$ in the bin. Then the expected regret due to the ill-behaved bins is bounded by

$$c'K^{-\beta} \left( \sum_{i \in J} \frac{N}{K} \right) \leq c'K^{-\beta}N \left( \frac{b}{K} + P \left( 0 < |f_1(t) - f_2(t)| \leq c'K^{-\beta} \right) \right)$$

$$\leq C'N \left( K^{\beta-1} + K^{\beta(1+\alpha)} \right)$$

$$\leq CNK^{-\beta(1+\alpha)}$$

where we have used the margin condition in the second to last inequality, and proposition 3 in the last inequality. $C$ is a positive constant.

For the well-behaved bins, the separation of the expectation functions, $|f_1(t) - f_2(t)| > cK^{-\beta}$, guarantees that the optimal strategy is constant within each bin (i.e. the functions do not cross). Without loss of generality, suppose $f_1$ is larger. Define the gaps, $\Delta_i$ as

$$\Delta_i = \inf_{B_i} f_1(t) - \sup_{B_i} f_2(t)$$

By our choice of the constant $c$ as at least $3L$, the Hölder continuity condition implies that the gaps $\Delta_i$ will be at least $LK^{-\beta}$. Lemma 4 provides an upper bound on the expected number of mistakes in any well-behaved bin of

$$\frac{8 \ln(N/K)}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

Our definition of the well-behaved bins and the Hölder continuity condition we have assumed imply that the expected regret per mistake is bounded by $3\Delta_i$. Thus, expected regret from well-behaved bins is bounded by

$$\sum_{i \in I} \left( (3 + \pi^2)\Delta_i + \frac{24\ln(N/K)}{\Delta_i} \right) \leq C \left( K + \sum_{i \in I} \frac{\ln(N/K)}{\Delta_i} \right)$$

Therefore, total expected regret for the dynamic policy is bounded by
for some positive constant $C$.

We now apply the margin condition to derive more information about the gaps $\Delta_i$, $i \in \mathcal{I}$, in order to obtain tighter bounds. Suppose that we have $K' \leq K$ of these bins and order the gaps $\Delta_1 \leq \Delta_2 \leq \ldots \leq \Delta_{K'}$. We note the margin condition ensures that

$$\lambda \{ t : 0 < |f_1(t) - f_2(t)| \leq \Delta_i + 2LK^{-\beta} \} \leq C(\Delta_i + 2LK^{-\beta})^\alpha \leq C' \Delta_i^\alpha$$

and the Hölder continuity condition combined with the monotonicity of the gaps implies that

$$\lambda \{ t : 0 < |f_1(t) - f_2(t)| \leq \Delta_i + 2LK^{-\beta} \} \geq \frac{i}{K}$$

Combining these two inequalities with the characterization of well-behaved bins shows that if $i \in \mathcal{I}$ we have

$$\Delta_i \geq C \left[ \max \left( \left( \frac{i}{K} \right)^{1/\alpha}, K^{-\beta} \right) \right]$$

This can be substituted to obtain a regret bound of

$$C \left( NK^{-\beta(1+\alpha)} + K + \sum_{i \in \mathcal{I}} \frac{\ln(N/K)}{\Delta_i} \right)$$

and by splitting the sum we can derive a more explicit bound:

$$\sum_{i=1}^{K'} \frac{\ln(N/K)}{\max \left( \left( \frac{i}{K} \right)^{1/\alpha}, K^{-\beta} \right)} \leq C \left[ K^{1-\alpha\beta+\beta} + \sum_{i=[K^{1-\alpha\beta}] + 1}^{K'} (i/K)^{-1/\alpha} \right] \ln(N/K)$$

The remaining sum can be bounded above by

$$CK \int_{K^{-\alpha\beta}}^{1} x^{-1/\alpha} \, dx$$
If $\alpha < 1$ this integral is bounded by $cK^{\beta(1-\alpha)}$. If $\alpha > 1$ the integral is bounded by $C \ln K$. Consequently, the entire term is always bounded by $CK^{1-\alpha \beta + \beta}$. Therefore, the total regret is bounded by

$$C \left[ NK^{-\beta(1+\alpha)} + K + K^{1-\alpha \beta + \beta} \ln(N/K) \right]$$

and the result follows from using the value of $K$ prescribed in the statement of the theorem. If $\alpha > 1$ the term $K$ dominates, if $\alpha \leq 1$ the other two terms dominate the asymptotic behavior of the regret.

This result shows that a relatively straightforward adaptation of an optimal policy for the static problem can achieve non-trivial regret growth in a dynamic setting. Furthermore, using Dynamic UCB Policy 1, we can expect to achieve the same asymptotic regret bounds for the multi-armed problem as we have in the two-arm case. So long as the margin condition is satisfied between the best arm and the second best arm, the proof of our theorem can be used to independently bound the regret due to each arm.

We should note that the regret bound attained in the theorem is clearly suboptimal in the case when $\alpha > 1$. The policy as presented has the advantage of allowing us to choose the number of bins, $K$, in ignorance of the value of $\alpha$, but we should recognize that knowledge of this parameter allows us to slightly modify our policy to do better in the case when $\alpha > 1$. Indeed, looking again at the bound on expected regret of

$$C \left[ NK^{-\beta(1+\alpha)} + K + K^{1-\alpha \beta + \beta} \ln(N/K) \right]$$

we see that if $\alpha > 1$ we can choose $K$ as

$$K = C_1 \left( \frac{N}{\ln N} \right)^{\frac{1}{\beta(1+\alpha)+1}}$$

to achieve a bound on asymptotic regret of

$$C_2 N \left( \frac{N}{\ln N} \right)^{-\frac{\beta(1+\alpha)}{\beta(1+\alpha)+1}}$$

which is an improvement over the theorem.

Discontinuities
We should also consider whether or not we can relax the assumptions in our theorem. A place where this is particularly desirable is the assumption of Hölder continuous reward expectation functions. If we relax this assumption by allowing the reward expectations to exhibit a finite number of jump discontinuities, we then have only to consider a second class of ill-behaved bins which contain these discontinuities. The number of such bins will be bounded by a constant, and the number of plays in these bins will thus be proportional to \( N/K \). Since we can make no real guarantees with regards to the regret per play, this means that the regret from these bins can be of order \( N/K \). If we add this term, we clearly lose a lot from our bound on the regret. We can balance the growth of \( K \) and \( N/K \) terms by selecting \( K \) proportional to \( \sqrt{N} \). If \( \alpha \geq \frac{1}{\beta} - 1 \), the \( K \) and \( N/K \) terms dominate, and if we know to expect jump discontinuities, we can achieve regret proportional to \( \sqrt{N} \). If \( \alpha < \frac{1}{\beta} - 1 \), then the other two terms dominate and we get a regret bound proportional to \( N^{0.5(1-\alpha\beta+\beta)} \ln(N) \).

**Theorem 6.** Consider a high frequency dynamic two-armed bandit whose reward expectation functions satisfy Hölder continuity—except possibly for a finite set of jump discontinuities—and the margin condition. Implementing Dynamic UCB Policy 1 with \( K = C_1 \sqrt{N} \), will yield expected regret bounded by
\[
C_2 \max \left( \sqrt{N}, N^{0.5(1-\alpha\beta+\beta)} \ln(N) \right),
\]
where \( C_1 \) and \( C_2 \) are positive constants.

We have shown that Dynamic UCB Policy 1 is quite robust in its ability to adapt to changing reward expectations. However, our present treatment of the problem only indirectly deals with the dynamic nature of the model. We are still estimating the functions \( f_i \) by taking simple averages which fail to account for continuously changing reward distributions within each bin. In the next few sections we will attempt to improve upon our performance by refining the manner in which we estimate the reward expectation for each machine. We begin laying the groundwork for our refinement by generalizing the approach used to derive a policy for the static model.

### 4.2 Generalized Upper Confidence Bound Policies

For the moment, we return to the static multi-armed bandit problem outlined in the introduction and recall the static allocation policy introduced in chapter 3 which achieved optimal regret growth. We now consider how our analysis of this policy must change if we replace the average, \( \bar{X}_{i,\theta_i(n)} \)—used as our estimate of the reward expectation—with an arbitrary weighted average of the past rewards. That is, our estimate of the reward expectation for machine \( i \) after \( n \) plays will now be equal to the dot product of the vector of realized rewards from machine \( i \) with an \( n \)-vector \( \theta_{i,n} \) satisfying the following properties:

1. \( \sum_{j=1}^{n} \theta_{i,n,j} = 1 \)
2. For all \( j, \theta_{i,n,j} \geq 0 \)
Static UCB Policy

- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ always play that machine which maximizes $X_{i,\#(n)} + c_{n,\#(n)}$, where $c_{t,s} = \sqrt{2 \ln(t)/s}$.

3. $\theta_{i,n,j} = 0$ if $I_j \neq i$

We will denote by $\hat{\mu}_{i,\#(n)}$ an estimator for the expectation $\mu_i$ after $n$ plays. This change in our estimator will necessitate a change in the way we approach the error term in an upper confidence bound. The new error term will be a function, $c(\theta_t)$, dependent on the vector of weights. Specifically, the error term will depend upon the bound we can place on the variance of our estimator:

$$E[(\hat{\mu}_{i,\#(n)} - E[\hat{\mu}_{i,\#(n)}])^2] \leq \|\theta_{t,n}\|^2$$

We now describe a general class of policies for the static problem and associated regret bounds.

Generalized Static UCB Policy

- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ always play that machine which maximizes $\hat{\mu}_{i,\#(n)} + c(\theta_{t,n})$, where $c(\theta_t) = \|\theta_t\|\sqrt{2 \ln(t)}$ and $\|\theta_t\| \leq \gamma \sqrt{\frac{1}{\#(t)}}$ for some constant $\gamma \geq 1$.

**Theorem 7.** A Generalized Static UCB Policy used on the static multi-armed bandit problem will achieve expected regret bounded by

$$\sum_{i \neq \ast} \left[ \frac{8\gamma^2 \ln(n)}{(\mu^* - \mu_i)} + (\mu^* - \mu_i) \left( 1 + \frac{\pi^2}{3} \right) \right]$$

**Proof.** Bounding the regret achieved by our generalized upper confidence bound policy after $n$ plays in the static problem will follow the same analysis used in
Chapter 3. We bound total expected regret by bounding $\mathbb{E}[\#_i(n)]$ for each machine. As before, we have the following bound involving a triple sum

$$\#_i(n) \leq t + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{l} \chi(\hat{\mu}_s^* + c(\theta_s^*) \leq \hat{\mu}_i,s_i + c(\theta_{i,t}))$$

where the estimators for each value of $s$ are computed using the first $s$ observations of the corresponding machine. For the condition $\hat{\mu}_i,s + c(\theta_{i,t}) \geq \hat{\mu}_s^* + c(\theta_s^*)$ to hold true, at least one of these statements must be true:

1. $\hat{\mu}_s^* \leq \mu^* - c(\theta_t^*)$,
2. $\hat{\mu}_{i,s} \geq \mu_i + c(\theta_{i,t})$, or
3. $\mu^* < \mu_i + 2c(\theta_{i,t})$.

Our definition of $c(\theta_t) = ||\theta_t||\sqrt{2\ln(t)}$ will allow us to bound the probabilities of the first two possibilities. Note that if vector $\theta_{i,t}$ is evenly weighted over the $\#_i(t)$ observations of machine $i$ that have been made as of time $t$, then

$$||\theta_{i,t}|| = \sqrt{\sum_{j=1}^{t} \theta_{i,t,j}^2}$$

$$= \sqrt{\#_i(t) \frac{1}{\#_i(t)^2}}$$

$$= \sqrt{\frac{1}{\#_i(t)}}$$

In this case, our choice of the confidence bound $c(\theta_t)$ is equivalent to that used in the static policy; our choice of the function $c(\theta_t)$ is a natural generalization of the static confidence bound.

Recall lemma 1 (Hoeffding’s inequality), which told us that if $S$ was the sum of a set of random variables, $X_i$, which were supported almost surely on the intervals $[a_i, b_i]$, then the following inequality holds

$$P(S - \mathbb{E}[S] \geq \delta) \leq e^{-\frac{\delta^2}{\sum_{i=1}^{n} (b_i - a_i)^2}}$$

We write our estimator at time $t$ after $s$ plays of machine $i$ as

$$\hat{\mu}_{i,s} = \sum_{j \leq i, l_j = i} \theta_{i,t,j} y_{i,j}$$
where $y_{i,j}$ is the realized reward from machine $i$ at time $j$. Using our assumptions on the reward distribution for the static problem, we know that the random variable $\theta_{i,t,j} y_{i,j}$ is supported almost surely on the interval $[0, \theta_{i,t,j}]$. Thus, by Hoeffding’s inequality

$$P(\hat{\mu}_{i,s} - \mu_i \geq \|\theta_{i,t}\| \sqrt{2 \ln(t)}) \leq e^{-\frac{4 \ln(t) \|\theta_{i,t}\|^{2}}{\sum_{j=1}^{t} \theta_{i,t,j}^{2}}} = e^{-4 \ln(t)} = \frac{1}{t^4}$$

This holds regardless of the ordering of our samples and regardless of the individual weights we place on the samples. Our weighting scheme can even change as time goes on, as long as the confidence bound at each time step is computed based upon the same vector of weights used to compute the estimator at that time. Therefore, the probability that $\hat{\mu}^*_{i,s} \leq \mu^* - c(\theta_{i,t})$ and the probability that $\hat{\mu}_{i,s} \geq \mu_i + c(\theta_{i,t})$ both contribute no more than a constant of $1 + \frac{\pi^2}{3}$ to our bound on $E[\#_i(n)]$.

We would like to again be able to choose $l$ so that $\mu^* < \mu_i + 2c(\theta_{i,t})$ cannot happen so long as $\#_i(t) \geq l$. In order to do this, however, we will need to use the additional assumption on our weight vectors. To achieve logarithmic regret growth, it must be the case that $\|\theta_{i,t}\| \leq \gamma \sqrt{\frac{1}{\pi_i(t)}}$ for some constant $\gamma \geq 1$. Note that the boundary case $\gamma = 1$ reduces to the static policy. With this final assumption on the vectors $\theta_{i,t}$, we can choose $l \geq 8\gamma^2 \ln(n)/(\mu^* - \mu_i)^2$ to guarantee that $\mu^* \geq \mu_i + 2c(\theta_{i,t})$. Thus, we can bound the expected number of plays of the suboptimal machine $i$ by

$$E[\#_i(n)] \leq \frac{8\gamma^2 \ln(n)}{(\mu^* - \mu_i)^2} + 1 + \frac{\pi^2}{3}$$

and total expected regret is then bounded by

$$\sum_{i \neq \ast} \left[\frac{8\gamma^2 \ln(n)}{(\mu^* - \mu_i)^2} + (\mu^* - \mu_i) \left(1 + \frac{\pi^2}{3}\right)\right]$$

We have now described a large class of allocation policies which achieve logarithmic regret growth for the static problem. Of course, in the case of the static problem there is no reason to use a policy with $\gamma > 1$ because there is no benefit to using non-constant weights: it simply increases the constant in our regret growth. The most important insight to be gained from examining this
general case is that achieving good regret bounds requires tight control over the variance of whatever estimator is used. The implication for the dynamic problem is this: we may be able to improve our allocation policies by reducing the bias of the estimator while retaining necessary bounds on the variance. We can thus view our problem as a classic bias-variance tradeoff. The next section of this chapter will deal with the problem of optimal estimation of the \( f \), given an arbitrary constraint on the variance.

4.3 Optimal Estimation

We begin dealing with the problem of estimation by considering the simple case of a one-armed bandit; we wish to estimate the expectation of a single machine as in the dynamic bandit problem described at the beginning of this chapter. The machine has expectation given by the Hölder continuous function \( f \), and this machine is played at every successive time step. We define an estimator at time \( \frac{t}{N} \) to be a function of the form

\[
\hat{f} \left( \frac{t}{N} \right) = \sum_{i=1}^{t} \theta_{t,i} y_i
\]

where \( y_i \) is the realization of this machine at time \( \frac{t}{N} \). Thus, our estimate of the function \( f \) is a weighted average of the past realizations, and we allow that the best choice of weights may change at each successive time step. This point is emphasized by the inclusion of the \( t \) subscript for the vector \( \theta_t \)

Our goal is to optimize this estimator by minimizing the bias relative to \( f \left( \frac{t+1}{N} \right) \) subject to an arbitrary variance constraint \( \| \theta_t \|^2 \leq V \). This will require us to construct a bound on the bias of our estimator. The bias is defined as

\[
\left| E \left[ \hat{f} \left( \frac{t}{N} \right) \right] - f \left( \frac{t+1}{N} \right) \right|
\]

To bound the bias we will employ our assumption that \( f \) is Hölder continuous with parameter \( \beta \leq 1 \) and constant \( L \). Thus we have
\[
\left| E \left[ \hat{f} \left( \frac{t}{N} \right) \right] - f \left( \frac{t+1}{N} \right) \right| = \left| \sum_{i=1}^{t} \theta_{t,i} E[y_i] - f \left( \frac{t+1}{N} \right) \right| \\
= \left| \sum_{i=1}^{t} \theta_{t,i} \left( f \left( \frac{i}{N} \right) - f \left( \frac{t+1}{N} \right) \right) \right| \\
\leq LN^{-\beta} \sum_{i=1}^{t} (t+1-i)^{\beta} \theta_{t,i}
\]

**Estimator Optimization Problem**

The problem of estimating the expectation for the one-armed dynamic bandit leads to a relatively straightforward convex optimization problem. We minimize the bound on the bias

\[
LN^{-\beta} \sum_{i=1}^{t} (t+1-i)^{\beta} \theta_{t,i}
\]

subject to the following set of constraints

1. \( \theta_{t,i} \geq 0 \) for each \( i, 1 \leq i \leq t \)
2. \( \sum_{i=1}^{t} \theta_{t,i} = 1 \)
3. \( \|\theta_t\|^2 \leq V \)

The Lagrangian for this problem is given by

\[
LN^{-\beta} \sum_{i=1}^{t} (t+1-i)^{\beta} + \lambda_0 \left( \|\theta_t\|^2 - V \right) - \sum_{i=1}^{t} \lambda_i \theta_{t,i} - \nu \left( \sum_{i=1}^{t} \theta_{t,i} - 1 \right)
\]

The Karush-Kuhn-Tucker conditions for this problem are

1. \( \theta_{t,i} \geq 0 \)
2. \( \|\theta_t\|^2 \leq \gamma^2 / t \)
3. \( \sum_{i=1}^{t} \theta_{t,i} = 1 \)
4. \( \lambda_i \geq 0 \) for each \( i, 0 \leq i \leq t \)
5. \( \lambda_i \theta_{t,i} = 0 \) for each \( i, 1 \leq i \leq t \)
6. $\lambda_0 \left( \|\theta_t\|^2 - V \right) = 0$

7. $LN^{-\beta}(t + 1 - i)^\beta + 2\lambda_0 \theta_{t,i} - \lambda_t - \nu = 0$ for each $i$, $1 \leq i \leq t$

Substituting condition 7 into condition 5 yields

$$\theta_{t,i} \left( LN^{-\beta}(t + 1 - i)^\beta + 2\lambda_0 \theta_{t,i} - \nu \right) = 0$$

$\nu < 0$ would imply $\theta_{t,i} = 0$ for all $i$, violating the third condition. Therefore we have $\nu \geq 0$ and for each $i$, $1 \leq i \leq t$ we have

$$\theta_{t,i} = \max \left( 0, \frac{\nu - LN^{-\beta}(t + 1 - i)^\beta}{2\lambda_0} \right)$$

The constraint on the variance is binding, so condition 6 in fact gives us the equation $\|\theta_t\|^2 - V = 0$, and this combined with condition 3 gives enough information to solve for $\nu$ and $\lambda_0$. Observe that the $\theta_{t,i}$ are monotonically increasing. This implies that there exists a non-negative integer $i^*$ such that if $i \leq i^*$ we have $\theta_{t,i} = 0$ and if $i > i^*$ then $\theta_{t,i} > 0$. Using this observation, we can substitute the expression for $\theta_{t,i}$ above into conditions 3 and 6 to derive the following two equations:

$$\frac{\sum_{i=i^*+1}^t \nu - LN^{-\beta}(t + 1 - i)^\beta}{2\lambda_0} = \frac{(t - i^*)\nu - \sum_{i=i^*+1}^t LN^{-\beta}(t + 1 - i)^\beta}{2\lambda_0} = 1$$

$$\frac{\sum_{i=i^*+1}^t \left( \nu - LN^{-\beta}(t + 1 - i)^\beta \right)^2}{4\lambda_0^2} = V$$

Squaring the first yields

$$1 = \frac{1}{4\lambda_0^2} (t - i^*)^2 \nu^2 - \frac{2}{4\lambda_0^2} (t - i^*)\nu \sum_{i=i^*+1}^t LN^{-\beta}(t + 1 - i)^\beta$$

$$+ \frac{1}{4\lambda_0^2} \left( \sum_{i=i^*+1}^t LN^{-\beta}(t + 1 - i)^\beta \right)^2$$

Expanding the squared binomial in the second and multiplying by $t - i^*$ gives us
\[ (t - i^*) V = \frac{1}{4\lambda_0} (t - i^*)^2 \nu^2 - \frac{2}{4\lambda_0} (t - i^*) \nu \sum_{i = i^* + 1}^{t} LN^{-\beta} (t + 1 - i)^\beta + \frac{1}{4\lambda_0^2} (t - i^*) \sum_{i = i^* + 1}^{t} (LN^{-\beta} (t + 1 - i)^\beta)^2 \]

Taking the difference and rearranging slightly leaves us with

\[ 4\lambda_0^2 = \frac{(t - i^*) \sum_{i = i^* + 1}^{t} (LN^{-\beta} (t + 1 - i)^\beta)^2 - \left( \sum_{i = i^* + 1}^{t} LN^{-\beta} (t + 1 - i)^\beta \right)^2}{(t - i^*) V - 1} \]

From here, we can solve for \( \nu \) by substituting into condition 3 as

\[ \nu = \frac{1}{t - i^*} \left( 2\lambda_0 + \sum_{i = i^* + 1}^{t} LN^{-\beta} (t + 1 - i)^\beta \right) \]

Unfortunately, there is not in general a closed form solution for the last remaining parameter, \( i^* \), and the generalization of this estimation approach to a multi-armed bandit significantly complicates this part of the problem. We presently turn our attention to generalizing the results of our estimator optimization to the multi-armed problem described at the outset of this chapter. We add the constraints that \( \theta_{j,t,i} = 0 \) if \( I_i \neq j \) (i.e. machine \( j \) was not played at time \( i/N \)). The manipulations performed so far in this section change very little and will not be reproduced. After adding these constraints, the estimator for machine \( j \) after the observation at time \( t/N \) can be described as follows:

\[ \theta_{j,t,i} = \max \left( 0, \nu - LN^{-\beta} (t + 1 - i)^\beta \right) \frac{2\lambda_0}{2\lambda_0} \]

if \( I_i = j \), and \( \theta_{j,t,i} = 0 \) otherwise

\[ 4\lambda_0^2 = \frac{\# \{ i : I_i = j, i > i^* \} \sum_{\{ i : i = j, i > i^* \} V - 1} (LN^{-\beta} (t + 1 - i)^\beta)^2 - \frac{1}{\# \{ i : I_i = j, i > i^* \} V - 1} \left( \sum_{\{ i : i = j, i > i^* \} LN^{-\beta} (t + 1 - i)^\beta \right)^2}{\frac{1}{\# \{ i : I_i = j, i > i^* \} V - 1} \left( \sum_{\{ i : i = j, i > i^* \} LN^{-\beta} (t + 1 - i)^\beta \right)^2} \]

\[ \nu = \frac{1}{\# \{ i : I_i = j, i > i^* \} \left( 2\lambda_0 + \sum_{\{ i : i = j, i > i^* \} LN^{-\beta} (t + 1 - i)^\beta \right)} \]

30
As before, \( \#_j(t) \) denotes the number of times machine \( j \) has been played during the first \( t \) time steps, and \( \#(S) \) is used to denote the number of elements in the set \( S \).

This estimation process suggests a number of ways to refine our first dynamic allocation policy which may improve performance. More importantly, conceptualizing the problem as a bias-variance tradeoff provides a framework through which we can evaluate alternative policies. The next section will consider several possible approaches to dynamic allocation informed by this work.

4.4 Refined Dynamic UCB Allocation Policies

The most obvious refinement suggested by our work on estimation is to modify Dynamic UCB Policy 1 by replacing the use of the static policy within each bin by the use of a generalized static policy. The weights for the generalized static policy could be selected in the manner outlined in section 4.3, and the parameter \( \gamma \)—representing the level of bias reduction desired—would need to be tuned for each application. Since generalized static policies have the same regret growth properties as the static policy introduced in chapter 3, the modified Dynamic policy would inherit the same regret bound we proved for Dynamic UCB Policy 1.

This modification allows a great deal of flexibility to adjust to particular applications, but it also has a number of weaknesses. The policy still has to discard information at regular intervals. Furthermore, implementing a generalized static policy with optimized weights is computationally intensive—generally, such an algorithm will require run time proportional to \( N^2 \). We will outline, without proving regret bounds, two alternative approaches to dynamic allocation in an attempt to address these weaknesses. Both policies adapt techniques that have already been outlined to reduce estimator bias, retain historical information, and enable simple implementations with run time proportional to \( N \). The success of these policies is assessed by the simulations presented in the next chapter.

Suppose we have an estimator for a function \( f \) at time \( t - 1 \) described by a vector of weights \( \theta_{t-1} \). If we make an observation of \( f \) at time \( t \), we can update our estimate by discounting the past by a factor \( \delta \). That is, the first \( t - 1 \) values in the vector \( \theta_t \) are given by \( \delta \theta_{t-1} \), and the weight given to the observation at time \( t \) is then \( 1 - \delta \) to preserve the property that weights sum to 1. Our choice of \( \delta \) can be guided by the variance constraint we imposed on generalized static UCB policies. In the case that an observation has been made of \( f \) at time \( t \) this constraint is

\[
\|\theta_t\| \leq \gamma \sqrt{\frac{1}{\#(t)}}
\]
If we choose $\delta$ to minimize bias subject to this constraint, then the constraint is binding and the optimal $\delta$ is given by

$$
\delta = 1 - \sqrt{\frac{\frac{\gamma^2}{\#(t)} - \|\theta_{t-1}\|^2 + \frac{\gamma^2\|\theta_{t-1}\|^2}{\#(t)}}{1 + \|\theta_{t-1}\|^2}}
$$

The estimator does not change if no observation is made at time $t$. The new entry of $\theta_i$ is filled with a zero. This process of discounting past observations when forming estimates will be central to both policies we now describe.

**Dynamic UCB Policy 2**

- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ always play that machine which maximizes $\hat{f}_{i,\#,i(n)} + c(\theta_{i,n})$, where $c(\theta_t) = \|\theta_t\|\sqrt{2\ln(t)}$, $\hat{f}_{i,\#,i(n)}$ is a weighted estimator for $f_i$ after $n$ plays computed using the discounting process described above, and $\theta_{i,n}$ is the associated vector of weights.

**Dynamic UCB Policy 3**

- Choose the number of bins, $K$, proportional to $(N/\ln N)^{1/(2\beta+1)}$
- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ in the interval $[j/K, (j + 1)/K]$, always play that machine which maximizes $\hat{f}_{i,\#,i,n} + c_{n_j,\#,i,j(n)}$, where $\hat{f}_{i,\#,i(n)}$ is a weighted estimator for $f_i$ after $n$ plays computed using the discounting process described above, $c_{t,s} = \sqrt{2\ln(t)/s}$, $n_j$ denotes the total number of plays so far in the interval $[j/K, (j + 1)/K]$, and $\#_{i,j}(n)$ denotes the number of plays so far of machine $i$ in the interval $[j/K, (j + 1)/K]$.

Dynamic UCB Policy 2 is essentially a generalized static policy implemented with the weighting scheme just described. Dynamic UCB Policy 3 is a hybrid policy modeled after Dynamic UCB Policy 1: we replace the bin average with a weighted estimator.
4.5 Non-Uniform Arrival Times

A notable weakness in the model as described up to this point is the assumption that decisions happen at a constant rate. When considering applications we often associate an allocation decision with the arrival of a potential customer, patient, etc., and in reality we cannot rely on such uniform arrivals. Fortunately, the policies described thus far do not lose any strength if we allow variation in the arrival process—provided there is a corresponding modification of the margin condition.

Consider the dynamic model introduced at the beginning of this chapter, except instead of considering arrivals that occur after every $1/N$ units of time, we consider arrivals given by a non-homogeneous Poisson process with parameter $\lambda(t) \geq 0$ such that

$$\frac{1}{N} \int_0^1 \lambda(t) \, dt = 1$$

The expected number of arrivals is thus still equal to $N$. Our previous work can be applied to achieve similar regret bounds for this more general bandit problem, but we will need to adjust the statement of the margin condition to ensure that ill-behaved bins are not concentrated around times with high arrival intensity. To facilitate this, we define an arrival density measure of a set $S$ on the interval $[0, 1]$ by

$$\mathcal{P}_\lambda(S) = \frac{1}{N} \int_S \lambda(t) \, dt$$

Take $\delta_0$, $C$, and $\alpha$ to be three positive constants. We say that a dynamic bandit problem with a non-homogeneous Poisson arrival process as described above satisfies the \textit{weighted margin condition} if for any two machines, $i$ and $j$, we have

$$\mathcal{P}_\lambda\{t : 0 < |f_i(t) - f_j(t)| \leq \delta\} \leq C\delta^\alpha$$

for any $t \in [0, 1]$ and any positive $\delta$ with $\delta < \delta_0$. Note that if $\lambda(t)$ is uniformly bounded, the margin condition implies the weighted margin condition. Conversely, if $\lambda(t)$ is bounded away from zero ($\lambda(t) \geq \delta$ for some $\delta > 0$), the weighted margin condition implies the margin condition.

Proposition 3 still accurately describes the restrictions on the values $\alpha$ and $\beta$ can hold simultaneously provided that the functions $f_i$ and $f_j$ intersect at a time $t$ such that $\lambda > 0$ in a neighborhood of $t$. If the reward expectations satisfy the weighted margin condition, we can apply Dynamic UCB Policy 1 to the bandit problem with non-uniform arrival times and we will achieve the same result.
on the regret as we did in section 4.1. In our proof we require the additional assumption that $\lambda(t)$ is bounded away from zero so that the original margin condition is satisfied, but this is not a restrictive assumption since intervals with zero arrival intensity do not add to our regret.

**Theorem 8.** Consider a high frequency dynamic two-armed bandit with a non-homogenous Poisson arrival process described by the function $\lambda(t)$, which is bounded away from zero. If the reward expectation functions satisfy H"older continuity and the weighted margin condition, then implementing Dynamic UCB Policy 1 with $K = C_1 \left( N/\ln N \right)^{1/(2\beta+1)}$ will yield expected regret bounded by

$$C_2 \frac{N}{\ln N} \left( N/\ln N \right)^{-\beta(1+\min(\alpha,1))}$$

where $C_1$ and $C_2$ are positive constants.

**Proof.** We again index successive bins $B_i, i = 1, 2, ..., K$ and characterize well-behaved bins by

$$\mathcal{I} = \{ i : \forall t \in B_i, |f_1(t) - f_2(t)| > cK^{-\beta} \}$$

for some $c \geq 3L$ where $L$ is the H"older continuity constant. If $i$ is not in $\mathcal{I}$ then the bin $B_i$ is said to be ill-behaved. H"older continuity allows us to find $c'$ such that $|f_1(t) - f_2(t)| \leq c'K^{-\beta}$ for any $t \in B_i$ where $i \notin \mathcal{I}$. Expected regret due to ill-behaved bins is therefore bounded by

$$c'K^{-\beta} N \mathcal{P}_{\lambda}\{ t : 0 < |f_1(t) - f_2(t)| \leq c'K^{-\beta} \} \leq CNK^{-\beta(1+\alpha)}$$

where we have used the weighted margin condition, and $C$ is a positive constant.

We can proceed exactly as before for the well-behaved bins to get expected regret from these bins bounded by

$$\sum_{i \in \mathcal{I}} \left( (3 + \pi^2)\Delta_i + \frac{24 \ln (N) \mathcal{P}_{\lambda}\{B_i}\}}{\Delta_i} \right) \leq C \left( K + \sum_{i \in \mathcal{I}} \frac{\ln(N)}{\Delta_i} \right)$$

so that total regret from Dynamic UCB Policy 1 is bounded by

$$C \left( NK^{-\beta(1+\alpha)} + K + \sum_{i \in \mathcal{I}} \frac{\ln(N)}{\Delta_i} \right)$$

for some positive constant $C$. The argument to bound the third term is again the same as before, and we obtain the following bound on regret.
Our choice of $K$ proves the theorem. 

This result assures us that Dynamic UCB Policy 1 will be able to maintain its performance in more general circumstances, widening the applicability of our model.

5 Simulations

Recall the main set of allocation policies discussed in chapters 3 and 4:

<table>
<thead>
<tr>
<th>Static UCB Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Initialize the policy by playing each machine once.</td>
</tr>
<tr>
<td>• After initialization, at time $n + 1$ always play that machine which maximizes $X_{t, #<em>i(n)} + c</em>{n, #<em>i(n)}$, where $c</em>{t,s} = \sqrt{2\ln(t)/s}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic UCB Policy 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Choose the number of bins, $K$, proportional to $(N/\ln N)^{1/(2\beta+1)}$</td>
</tr>
<tr>
<td>• Begin by implementing the Static UCB Policy</td>
</tr>
<tr>
<td>• At the beginning of each time interval $[j/K, (j + 1)/K]$, $j = 1, 2, 3, ..., K - 1$, reset by discarding all historical information and reinitialize the Static UCB Policy as though facing the start of a new bandit problem</td>
</tr>
</tbody>
</table>

We studied the empirical behavior of Dynamic UCB Policies 1, 2, and 3 in three principal manners through simulations in MATLAB. First, the policies were compared against each other using a series of two-armed problems ranging
Dynamic UCB Policy 2

- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ always play that machine which maximizes $\hat{f}_i,\#_i(n) + c(\theta_{i,n})$, where $c(\theta_t) = \|\theta_t\| \sqrt{2\ln(t)}$, $\hat{f}_i,\#_i(n)$ is a weighted estimator for $f_i$ after $n$ plays computed using the discounting process described in section 4.4, and $\theta_{i,n}$ is the associated vector of weights.

Dynamic UCB Policy 3

- Choose the number of bins, $K$, proportional to $(N/\ln N)^{1/(2\beta+1)}$.
- Initialize the policy by playing each machine once.
- After initialization, at time $n + 1$ in the interval $[j/K, (j + 1)/K]$, always play that machine which maximizes $\hat{f}_i,\#_i(n) + c_{n_j,\#_i,n}(n)$, where $\hat{f}_i,\#_i(n)$ is a weighted estimator for $f_i$ after $n$ plays computed using the discounting process described in section 4.4, $c_{t,s} = \sqrt{2\ln(t)/s}$, $n_j$ denotes the total number of plays so far in the interval $[j/K, (j + 1)/K]$, and $\#_i,n(n)$ denotes the number of plays so far of machine $i$ in the interval $[j/K, (j + 1)/K]$.

from quite simple to very difficult. The dynamic policies were also compared against the static allocation policy from chapter 3. Second, the sensitivity of each of the dynamic policies to various levels of noise was examined. Finally, experiments were performed with Dynamic UCB Policy 1 on the multi-arm case to compare against the regret growth observed in the two-arm problems.

Preliminary experiments were performed with policy 1 in order to roughly tune the parameter used as the leading constant for the number of bins $K$. Policy 1 is very sensitive to this choice of leading constant. This choice induces a tradeoff between the benefit from having a faster response to crossings and the benefit from less frequently having to re-initialize the strategy. The faster response benefit becomes less important as $N$ increases, motivating the use of a lower constant, but for some practical applications, as well as for some of our experiments with high numbers of crossings, a higher leading constant can substantially improve performance in lower ranges of $N$. We have found that
using $K = 0.2(N/\ln N)^{1/(2\beta+1)}$ bins seems to work well for most of the expected reward functions used in our experiments. To help illustrate the impact of the choice of leading constant, in the comparison experiments policy 1 is used with both $K = 0.2(N/\ln N)^{1/(2\beta+1)}$ bins and $K = 0.6(N/\ln N)^{1/(2\beta+1)}$ bins. Policies 2 and 3 have been similarly tuned. In both we use $\gamma = \sqrt{15}$, and in policy 3 we use $K = 0.2(N/\ln N)^{1/(2\beta+1)}$ bins.

5.1 Comparison Between Policies

The policies were compared against each other using a set of eight pairs of expected reward functions designed to test our policies in a wide range of possible scenarios. In each experiment, rewards were drawn from a Bernoulli distribution so that each realization was equal to either zero or one. The problems range from very simple single crossing problems, to problems with smooth oscillating rewards, to problems with rapidly changing rewards illustrating cases with $\beta = 1/3$. For each problem, we looked at two performance metrics: an estimation of the expected regret for the given scenario, and the percentage of suboptimal arm pulls. The estimates at each value of $N$ are the result of performing 100 experiments and averaging all observations. In appendix A we illustrate each of the eight problems—labeled problem 1 through problem 8—and for each policy we plot the natural log of the regret against $\ln N$ as well as the percentage of suboptimal decisions against $\ln N$. Box plots created from 100 experiments using $N = 100000$ are also provided to illustrate the variability in the performance of each policy.

In the majority of the problems we looked at, the static policy is vastly outmatched by the dynamic policies. All three dynamic policies achieve regret which is consistently sub-linear in $N$, while the static policy often approaches linear regret growth (see problems 2, 3, and 5 as examples). Dynamic UCB Policy 1 provides the most consistent performance across our set of problems. Dynamic UCB Policies 2 and 3 perform very well in some cases, but are prone to having wide variability in results from trial to trial. Policy 2 is especially erratic (see problems 3 and 7). Our simulations suggest that the unmodified Dynamic UCB Policy 1 would perform best among these options in a majority of applications of our model.

5.2 Sensitivity to Noise

We examined the sensitivity of Dynamic UCB policies 1, 2, and 3 to noise by repeating experiments for problems 3, 5, and 7 from the comparisons section using three different levels of noise: a high level of noise given by the same Bernoulli distribution as before, a medium level of noise in which reward realizations differ from expectations by no more than 0.1, and a low level of noise in which we make exact observations of the reward expectation functions. In
appendix B, plots of the natural log of the regret against ln N for each policy show performance with respect to all three levels of noise, and a box plot constructed from 100 trial experiments using N = 100000 is also shown for each policy. The box plot displays only the medium and high noise levels, as the low noise level produces no deviation from the expected regret. Policies 1 and 3 both used $K = 0.2(N/\ln N)^{1/(2\varepsilon+1)}$, and policies 2 and 3 use $\gamma = \sqrt{15}$.

Policy 2 is clearly the most sensitive to changes in the level of noise. With the exception of problem 5, Policy 2 exhibited a dramatic decrease in expected regret as the level of noise was lowered, and an extremely significant reduction in the variation from the mean regret between experiments. Policies 1 and 3 are both relatively stable in performance across all three of the problems. Reducing the level of noise has no significant impact on expected regret—it actually raised the expected regret in some cases.

5.3 Multiple Arms

Dynamic UCB Policy 1 was tested on three distinct problems—labeled problem 1M through 3M in appendix C—involving three separate arms. As before, an average over 100 trials is used to estimate the expected regret. These experiments were conducted with $K = 0.6(N/\ln N)^{1/(2\varepsilon+1)}$. Dynamic UCB Policy 1 exhibited similar regret growth to what was achieved in the two-armed case for each of these instances, though with higher regret growth constants. These experiments confirm what we would have expected based upon our work in section 4.1. A graph of the natural log of the regret against ln N is presented in appendix C.

Appendices

Descriptions of each of the policies used in our simulations are provided at the start of chapter 5. Refer to chapters 3 and 4 for details on implementation and regret bounds for the policies. In the graphs in the comparisons section, UCB 1 - 0.2 refers to Dynamic UCB Policy 1 implemented using $K = 0.2(N/\ln N)^{1/(2\varepsilon+1)}$ bins, while UCB 1 - 0.6 refers to Dynamic UCB Policy 1 implemented using $K = 0.6(N/\ln N)^{1/(2\varepsilon+1)}$ bins.

A Comparisons
Problem 1

\[ f_1(t) = 0.5 \]
\[ f_2(t) = t \]
<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.5672</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.6247</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.5221</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.7136</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.4918</td>
</tr>
</tbody>
</table>
Problem 2

\[ f_1(t) = 0.5 \]

\[ f_2(t) = -4(t - 0.5)^3 + 0.5 \]
### Problem 2 - Log Regret vs. Log N

<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.9795</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.6997</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.5662</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.7986</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.7084</td>
</tr>
</tbody>
</table>
Problem 3

\[ f_1(t) = 0.5t + 0.25 \]
\[ f_2(t) = -0.25 \sin(4\pi t) + 0.5 \]
<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>1.0861</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.7184</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.8031</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.9251</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.8735</td>
</tr>
</tbody>
</table>
Problem 4

\[ f_1(t) = \begin{cases} 
4.5(t - 1/3)^2 + 0.5 & \text{if } t \leq 1/3; \\
0.5 & \text{if } 1/3 < t < 3/4; \\
(t - 0.75)^2 + 0.5 & \text{if } t \geq 3/4.
\end{cases} \]

\[ f_2(t) = \begin{cases} 
0.75 & \text{if } t \leq 1/6; \\
-54(t - 1/3)^3 + 0.5 & \text{if } 1/6 < t < 1/3; \\
0.1 \sin(10\pi(t - 1/3)) + 0.5 & \text{if } t \geq 1/3.
\end{cases} \]
Policy & Log-log trendline slope

<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.9480</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.8208</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.9071</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.8885</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.8167</td>
</tr>
</tbody>
</table>
Problem 5

\[ f_1(t) = 0.1 \sin(10\pi t) + 0.85 - 0.7t \]

\[ f_2(t) = 0.2 \sin(2\pi t) + 0.5 \]
### Problem 5 - Log Regret vs. Log N

<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.9868</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.7648</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.7670</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.7739</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.7562</td>
</tr>
</tbody>
</table>
Problem 6

\[ f_1(t) = 0.5 \]
\[ f_2(t) = 0.1 \sin(100t^2) + 0.5 \]
<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>0.9296</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.8682</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.8795</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.9512</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.9625</td>
</tr>
</tbody>
</table>
Problem 7

\[ f_1(t) = 0.5 \]

\[ f_2(t) = \begin{cases} 
(−1)^{i+1}|(2i - 1)/20 - t|^{1/3} + 0.5 & \text{ if } (i - 1)/10 \leq t < (2i - 1)/20; \\
(−1)^i|(2i - 1)/20 - t|^{1/3} + 0.5 & \text{ if } (2i - 1)/20 \leq t < i/10.
\end{cases} \]

for \( i = 1, 2, ..., 10 \)
<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>1.0315</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.7927</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.7150</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.9215</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.8061</td>
</tr>
</tbody>
</table>

Problem 7 - Log Regret vs. Log N

Problem 7 - Mistake % vs Log N
**Problem 8**

\[ f_1(t) = \begin{cases} 
0.5(-1)^i[(2i-1)/20 - t]^{1/3} + 0.4 + 0.2t & \text{if } (i-1)/10 \leq t < (2i-1)/20; \\
0.5(-1)^{i+1}[(2i-1)/20 - t]^{1/3} + 0.4 + 0.2t & \text{if } (2i-1)/20 \leq t < i/10.
\end{cases} \]

\[ f_2(t) = \begin{cases} 
(-1)^i[2i-1)/20 - t]^{1/3} + 0.5 & \text{if } (i-1)/10 \leq t < (2i-1)/20; \\
(-1)^i[(2i-1)/20 - t]^{1/3} + 0.5 & \text{if } (2i-1)/20 \leq t < i/10.
\end{cases} \]

for \( i = 1, 2, \ldots, 10 \)
<table>
<thead>
<tr>
<th>Policy</th>
<th>Log-log trendline slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>1.0100</td>
</tr>
<tr>
<td>UCB 1 - 0.6</td>
<td>0.7378</td>
</tr>
<tr>
<td>UCB 1 - 0.2</td>
<td>0.5590</td>
</tr>
<tr>
<td>UCB 2</td>
<td>0.6583</td>
</tr>
<tr>
<td>UCB 3</td>
<td>0.7927</td>
</tr>
</tbody>
</table>
B  Sensitivity to Noise

Problem 3
See appendix A for problem specification and illustration.
Problem 5

See appendix A for problem specification and illustration.
Problem 7
See appendix A for problem specification and illustration.
\section*{C Multiple Arms}

\textbf{Problem 1M}

\[ f_1(t) = 0.5t + 0.25 \]
\[ f_2(t) = -0.25 \sin(4\pi t) + 0.5 \]
\[ f_3(t) = -0.25 \sin(4\pi t) + 0.5 \]

\textbf{Problem 2M}

\[ f_1(t) = 0.5t + 0.25 \]
\[ f_2(t) = -0.25 \sin(4\pi t) + 0.5 \]
\[ f_3(t) = 0.5 \]
**Problem 3m**

\[ f_1(t) = 0.1 \sin(10\pi t) + 0.85 - 0.7t \]

\[ f_2(t) = -0.2 \sin(2\pi t) + 0.5 \]

\[ f_3(t) = 0.5t + 0.25 \]
References


