Journal of Machine Learning Research volume (2012) 2069-2106 Submitted 7/11; Revised 5/12; Published 6/12

# Optimistic Bayesian Sampling in Contextual-Bandit Problems

**Benedict C. May** BEN.MAY@BRIS.AC.UK
*School of Mathematics*
*University of Bristol*
*Bristol, BS8 1TW, United Kingdom*

**Nathan Korda** KORDA@MATHS.OX.AC.UK
**Anthony Lee** LEE@STATS.OX.AC.UK
*Oxford-Man Institute*
*University of Oxford*
*Eagle House, Walton Well Road*
*Oxford, OX2 6ED, United Kingdom*

**David S. Leslie** DAVID.LESLIE@BRIS.AC.UK
*School of Mathematics*
*University of Bristol*
*Bristol, BS8 1TW, United Kingdom*

**Editor:** Nicolò Cesa-Bianchi

## Abstract

In sequential decision problems in an unknown environment, the decision maker often faces a dilemma over whether to explore to discover more about the environment, or to exploit current knowledge. We address the exploration-exploitation dilemma in a general setting encompassing both standard and contextualised bandit problems. The contextual bandit problem has recently resurfaced in attempts to maximise click-through rates in web based applications, a task with significant commercial interest.

In this article we consider an approach of Thompson (1933) which makes use of samples from the posterior distributions for the instantaneous value of each action. We extend the approach by introducing a new algorithm, Optimistic Bayesian Sampling (OBS), in which the probability of playing an action increases with the uncertainty in the estimate of the action value. This results in better directed exploratory behaviour.

We prove that, under unrestrictive assumptions, both approaches result in optimal behaviour with respect to the average reward criterion of Yang and Zhu (2002). We implement OBS and measure its performance in simulated Bernoulli bandit and linear regression domains, and also when tested with the task of personalised news article recommendation on a Yahoo! Front Page Today Module data set. We find that OBS performs competitively when compared to recently proposed benchmark algorithms and outperforms Thompson's method throughout.

**Keywords:** multi-armed bandits, contextual bandits, exploration-exploitation, sequential allocation, Thompson sampling

## 1. Introduction

In sequential decision problems in an unknown environment, the decision maker often faces a dilemma over whether to explore to discover more about the environment, or to exploit current knowledge. We address this exploration-exploitation dilemma in a general setting encompass-

©2012 Benedict C. May, Nathan Korda, Anthony Lee and David S. Leslie.

ing both standard bandit problems (Gittins, 1979; Sutton and Barto, 1998; Auer et al., 2002) and contextual-bandit problems (Graepel et al., 2010; Li et al., 2010; Auer, 2002; Yang and Zhu, 2002). This dilemma has traditionally been solved using either ad hoc approaches like $\varepsilon$-greedy or softmax action selection (Sutton and Barto, 1998, Chapter 2) or computationally demanding lookahead approaches such as Gittins indices (Gittins, 1979) which provably satisfy an optimality criterion with respect to cumulative discounted reward. However, the lookahead approaches become intractable in all but the simplest settings and the ad hoc approaches are generally perceived to over-explore, despite providing provably optimal long term average reward.

In recent years, Upper Confidence Bound (UCB) methods have become popular (Lai and Robbins, 1985; Kaelbling, 1994; Agrawal, 1995; Auer et al., 2002), due to their low computational cost, ease of implementation and provable optimality with respect to the rate of regret accumulation.

In this article we consider an approach of Thompson (1933) which uses posterior distributions for the instantaneous value of each action to determine a probability distribution over the available actions. Thompson considered only Bernoulli bandits, but in general the approach is to sample a value from the posterior distribution of the expected reward of each action, then select the action with the highest sample from the posterior. Since in our generalised bandit setting the samples are conditioned on the regressor, we label this technique as Local Thompson Sampling (LTS). The technique is used by Microsoft in selecting adverts to display during web searches (Graepel et al., 2010), although no theoretical analysis of Thompson sampling in contextual bandit problems has been carried out.

When these posterior samples are represented as a sum of exploitative value and exploratory value, it becomes clear that LTS results in potentially negative exploratory values. This motivates a new algorithm, Optimistic Bayesian Sampling (OBS), which is based on the LTS algorithm, which is modified by replacing negative exploratory value with a zero value.

We prove that, under unrestrictive assumptions, both approaches result in optimal behaviour in the long term consistency sense described by Yang and Zhu (2002). These proofs use elementary and coupling techniques.

We also implement LTS and OBS and measure their performance in simulated Bernoulli bandit and linear regression domains, and also when tested with the task of personalised news article recommendation on the the Yahoo! Front Page Today Module User Click Log Data Set (Yahoo! Academic Relations, 2011). We find that LTS displays competitive performance, a view shared by Chapelle and Li (2011), and also that OBS outperforms LTS throughout.

### 1.1 Problem Formulation

An agent is faced with a contextual bandit problem as considered by Yang and Zhu (2002). The process runs for an infinite sequence of time steps, $t \in \mathcal{T} = \{1,2,\ldots\}$. At each time step, $t$, a regressor, $x_t \in \mathcal{X}$, is observed. An action choice, $a_t \in \mathcal{A}$, $\mathcal{A} = \{1,\ldots,A\}, A < \infty$, is made and a reward $r_t \in \mathbb{R}$ is received.

The contextual bandit framework considered assumes that reward can be expressed as

$$r_t = f_{a_t}(x_t) + z_{t,a_t}$$

where the $z_{t,a}$ are zero mean random variables with unknown distributions and $f_a : \mathcal{X} \to \mathbb{R}$ is an unknown continuous function of the regressor specific to action $a$. The stream of regressors $x_t$ is assumed not to be influenced by the actions or the rewards, and for simplicity we assume that these

are drawn independently from some fixed distribution on $\mathcal{X}$.[1] For our actions to be comparable, we assume that $\forall a \in \mathcal{A}, \forall t \in \mathcal{T}, \forall x \in \mathcal{X}$, $f_a(x) + z_{t,a}$ is supported on the same set, $\mathcal{S}$. Furthermore to avoid boundary cases we assume that $\forall a \in \mathcal{A}$

$$\sup_{x \in \mathcal{X}} f_a(x) < \sup \mathcal{S}. \tag{1}$$

In situations where the $z_{t,a}$ have unbounded support, $\mathcal{S} = \mathbb{R}$, and (1) is vacuous if $\mathcal{X}$ is compact. The condition is meaningful in situations where $\mathcal{S}$ is compact, such as if rewards are in $\{0,1\}$.

**Definition 1** *The optimal expected reward function, $f^* : \mathcal{X} \to \mathbb{R}$, is defined by*

$$f^*(x) = \max_{a \in \mathcal{A}} f_a(x).$$

A minimal requirement for any sensible bandit algorithm is the average reward convergence criterion of Yang and Zhu (2002), which identifies whether a sequence of actions receives, asymptotically, rewards that achieve this optimal expected reward. Hence the main theoretical aim in this article is to prove under mild assumptions that LTS and OBS constructs a sequence of actions such that

$$\frac{\sum_{s=1}^{t} f_{a_s}(x_s)}{\sum_{s=1}^{t} f^*(x_s)} \xrightarrow{\text{a.s.}} 1 \text{ as } t \to \infty. \tag{2}$$

The choice of action $a_t$ is based on the current and past regressors, $\{x_1, \ldots, x_t\}$, past action choices, $\{a_1, \ldots, a_{t-1}\}$, and past rewards, $\{r_1, \ldots, r_{t-1}\}$. Denote $\tilde{I}_1 = \emptyset$ and, for all times $\{t \in \mathcal{T} : t \geq 2\}$, denote

$$\tilde{I}_t = (x_1, \ldots, x_{t-1}, r_1, \ldots, r_{t-1}, a_1, \ldots, a_{t-1}).$$

Furthermore denote all of the prior information available as $I_0$ and also all the information available at time $t$ as $I_t$ $(= I_0 \cup \tilde{I}_t)$.

**Definition 2** *The policy, $\left(\pi_t(\cdot)\right)_{t \in \mathcal{T}}$, is a sequence of conditional probability mass functions where $\pi_t(a) = \mathbb{P}(a_t = a | I_t, x_t)$. At each time step $t$, the policy maps $I_t$ and $x_t$ to a probability mass function giving the probability of each action being selected.*

The policy is constructed in advance of the process, using only $I_0$, and is the function used to map $I_t$ and $x_t$ to action selection probabilities for each of the actions.

Note also that, under a Bayesian approach, the information sets $I_t$ result in posterior distributions for quantities of potential interest. In particular $I_0$ defines the assumed functional forms of the $f_a$, and a prior distribution over the assumed space of functions, which is then updated as information is received, resulting in a Bayesian regression procedure for estimating the reward functions $f_a$, and hence a posterior distribution and expectation of $f_a(x_t)$ conditional on the information set $I_t \cup \{x_t\}$.

We do not however formulate an exact probability model of how regressors are sampled, rewards are drawn and inference is carried out. Instead we rely on Assumptions 1–5 placed on the Bayesian regression framework, given in Section 3, that will be satisfied by standard models for the $x_t$, $r_t$ and prior information $I_0$. In particular, randomness resulting from the regressor and reward sequences are controlled through these assumptions, whereas our proofs control the randomness due to the

1. Note that this assumption of iid sampling from $\mathcal{X}$ is only used in the latter part of the proof of Theorem 1. In fact an ergodicity condition on the convergence of sample averages would suffice, but would increase the notational complexity of the proofs.

action selection method. A useful framework to keep in mind is one in which regressors are drawn independently from a distribution on a compact Euclidean space $\mathcal{X}$, each $z_{t,a}$ is a Gaussian random variable independent of all other random variables, and the prior information $I_0$ includes that each $f_a$ is a linear function, and a prior distribution over the parameters of these functions; we revisit this model in Section 4.2 to demonstrate how this framework does indeed ensure that all the Assumptions are satisfied. However much more general frameworks will also result in our Assumptions being satisfied, and restricting to a particular probability model at this point will unnecessarily restrict the analysis.

### 1.2 Algorithm Motivation

The choice of algorithm presented in this article is motivated by both infinite and finite time considerations. The first subsection of this section describes desirable infinite time properties for an algorithm that are of importance in proving optimality condition (2). The second subsection describes, in a heuristic manner, desirable finite time properties to help understanding of the motivation behind our choice of algorithm, as opposed to the many other algorithms that also satisfy the infinite time requirements.

#### 1.2.1 INFINITE TIME CONSIDERATIONS

In conventional interpretations of similar problems (Littman, 1996; Singh et al., 2000; Sutton and Barto, 1998), there are two major aspects of generating a policy. The first is developing an evaluation scheme and the second an action selection scheme.

So that the agent can evaluate actions, a regression procedure is used to map the current regressor and the history $I_t$ to value estimates for the actions. Denote the agent's estimated value of action $a$ at time $t$ when regressor $x$ is presented as $\hat{f}_{t,a}(x)$. Since $\hat{f}_{t,a}$ is intended to be an estimate of $f_a$, it is desirable that the evaluation procedure is consistent, that is, $\forall a \in \mathcal{A}, \forall x \in \mathcal{X}$, $\hat{f}_{t,a}(x) - f_a(x)$ converges in some sense to 0 as $n_{t,a} \to \infty$, where $n_{t,a}$ is the number of times action $a$ has been selected up to time $t$. Clearly such convergence will depend on the sequence of regressor values presented. However consistency of evaluation is not the focus of this work, so will be assumed where necessary and the evaluation procedure used for all algorithms compared in the numerical experiments in §4 will be the same. The main focus of this work is on the action selection side of the problem.

Once action value estimates are available, the agent must use an action selection scheme to decide which action to play. So that the consistency of estimation is achieved, it is necessary that the action selection ensures that every action is selected infinitely often. In this work, we consider algorithms generating randomised policies as a way of ensuring infinite exploration is achieved.

In addition to consistent evaluation and infinite exploration, it is also necessary to exploit the obtained information. Hence the action selection method should be greedy in the limit, that is, the policy $\pi_t$ is designed such that

$$\sum_{a \in \operatorname{argmax}_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t)} \pi_t(a) \to 1 \text{ as } t \to \infty.$$

These considerations result in the consideration of GLIE (greedy in the limit with infinite exploration) policies, for which action selection is greedy in the limit and also guarantees infinite

exploration (Singh et al., 2000). We combine a GLIE policy with consistent evaluation to achieve criterion (2).

### 1.2.2 FINITE TIME CONSIDERATIONS

As well as convergence criterion (2), our choice of algorithm is also motivated by informal finite time considerations, since many algorithms for which (2) holds are perceived to explore more than is desirable. We note that formal optimality criteria are available, such as expected cumulative discounted reward (Gittins, 1979) and rate of regret accumulation (Auer et al., 2002). However an analysis of Thompson sampling under these criteria has proved elusive, and our heuristic approach inspires a modification of Thompson sampling which compares favourably in numerical experiments (see Section 4). In this section, we discuss the short term heuristics.

In particular, consider the methodology of evaluating both an exploitative value estimate and an 'exploratory bonus' at each time step for each action, and then acting greedily based on the sums of exploitative and exploratory values (Meuleau and Bourgine, 1999). An action's exploitative value estimate corresponds to the expected immediate reward (i.e., expected reward for the current timestep) from selecting the action, given information obtained so far, and therefore the posterior expectation of expected immediate reward is the appropriate exploitative action value estimate.

**Definition 3** *Let $p_a(\cdot \mid I_t, x_t)$ denote the posterior distribution of $f_a(x_t)$ given $I_t$ and $x_t$, and let $Q^{Th}_{t,a}$ be a random variable with distribution $p_a(\cdot \mid I_t, x_t)$. The exploitative value, $\hat{f}_{t,a}(x_t)$, of action $a$ at time $t$ is defined by*

$$\hat{f}_{t,a}(x_t) = \mathbb{E}(Q^{Th}_{t,a} \mid I_t, x_t).$$

Thompson (1933) suggests selecting action $a_t$ with probability equal to the probability that $a_t$ is optimal, given $I_t$ (there is no regressor in Thompson's framework). This principle has recently been used by Graepel et al. (2010), who implement the scheme by sampling, for each $a$, $Q^{\mathrm{Th}}_{t,a}$ from the posterior distribution $p_a(\cdot \mid I_t, x_t)$ and selecting an action that maximises $Q^{\mathrm{Th}}_{t,a}$. This corresponds to using an exploratory value $\tilde{f}^{\mathrm{Th}}_{t,a}(x_t) := Q^{\mathrm{Th}}_{t,a} - \hat{f}_{t,a}(x_t)$ which is sampled from the posterior distribution of the error in the exploitative action value estimate at the current regressor. We name this scheme Local Thompson Sampling (LTS), where 'local' makes reference to the fact that action selection probabilities are the probabilities that each action is optimal at the current regressor. Under mild assumptions on the posterior expectation and error distribution approximations used, one can show that Local Thompson Sampling guarantees that convergence criterion (2) holds (see Theorem 1).

However the exploratory value $\tilde{f}^{\mathrm{Th}}_{t,a}(x_t)$ under LTS has zero conditional expectation given $I_t$ and $x_t$ (by Definition 3) and can take negative values. Both of these properties are undesirable if one assumes that information is useful for the future. One consequence of this is that, in regular situations, the probability of selecting an action $\hat{a}^*_t \in \operatorname{argmax}_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t)$ decreases as the posterior variance of $f_{\hat{a}^*_t}(x_t) - \hat{f}_{t,\hat{a}^*_t}(x_t)$ increases, that is, if the estimate for an action with the highest exploitative value has a lot of uncertainty then it is less likely to be played than if the estimate had little uncertainty.

To counteract this feature of LTS, we introduce a new procedure, Optimistic Bayesian Sampling (OBS) in which the exploratory value is given by

$$\tilde{f}_{t,a}(x_t) = \max(0, \tilde{f}^{\mathrm{Th}}_{t,a}(x_t) - \hat{f}_{t,a}(x_t)).$$

This exploratory value has positive conditional expectation given $I_t$ and $x_t$ and cannot take negative values. The exploratory bonus results in increased selection probabilities for uncertain actions, a

desirable improvement when compared to LTS. In §3, we show that OBS satisfies the convergence criterion (2) under mild assumptions. Furthermore, simulations described in §4 indicate that the OBS algorithm does indeed outperform LTS, confirming the intuition above.

### 1.3 Related Work

There are three broad classes of exploration approach: undirected, myopic and belief-lookahead (Asmuth et al., 2009). In undirected exploration, the action selection distribution depends only on the values of the exploitative action value estimates. Examples of undirected exploration include ε-greedy and softmax action selection (see Chapter 2 of Sutton and Barto, 1998). In general, the short term performance of undirected methods is restricted by the fact that estimate uncertainty is not considered.

At the other end of the spectrum, in belief-lookahead methods, such as those suggested by Gittins (1979), a fully Bayesian approach is incorporated in which the action yielding the highest expected cumulative reward over the remainder of the process is selected,[2] thereby considering exploitative and exploratory value both directly and simultaneously and providing the optimal decision rule according to the specific criterion of maximising expected cumulative discounted reward. According to Wang et al. (2005),"in all but trivial circumstances, there is no hope of exactly following an optimal action selection strategy". Furthermore, even when it is possible to evaluate the optimal decision rule, "the optimal solutions are typically hard to compute, rely on artificial discount factors and fail to generalise to realistic reward distributions" (Scott, 2010). There is also the issue of 'incomplete learning'; Brezzi and Lai (2000) showed that, for standard bandit problems, Gittins' index rule samples only one action infinitely often and that this action is sub-optimal with positive probability. If the modelling assumptions and posterior approximations used are accurate, then this is a price worth paying in order to maximise expected cumulative discounted reward. However, if the posterior approximation method admits a significant error, then it may be that a too heavy reliance is placed on early observations. For these reasons, Gittins-type rules are rarely useful in practice.

In myopic methods, the uncertainty of action value estimates is taken into account, although the impact of action selections on future rewards is not considered directly. The exploratory component of myopic methods aims to reduce the uncertainty at the current regressor without explicitly considering future reward. By reducing uncertainty at each point presented as a regressor, uncertainty is reduced globally 'in the right places' without considering the regressor distribution. Myopic action selection can be efficient, easy to implement and computationally cheap. The LTS and OBS methods presented in this paper are myopic methods. The other main class of myopic methods are the upper confidence bound methods, which are now popular in standard and contextual bandit applications, and in some settings can be proved to satisfy an optimality criterion with respect to the rate of accumulation of regret (for an overview, and definitions of various notions of regret, see Cesa-Bianchi and Lugosi, 2006).

Inspired by the work of Lai and Robbins (1985) and Agrawal (1995), Auer et al. (2002) proposed a myopic algorithm, UCB1, for application in standard bandit problems. The exploratory value at time $t$ for action $a$, which we denote $\tilde{f}_{t,a}$, takes the simple form

$$\tilde{f}_{t,a} = \sqrt{\frac{2\log(t-1)}{n_{t,a}}}.$$

2. Note that this is only meaningful in the case of discounted rewards or if the time sequence is finite.

Infinite exploration is guaranteed by the method, since the exploratory value grows in periods in which the associated action is not selected. Moreover, Auer et al. (2002) prove that the expected finite-time regret is logarithmically bounded for bounded reward distributions, matching the (asymptotically) optimal rate derived by Lai and Robbins (1985) uniformly over time. Auer et al. (2002) also propose a variant of UCB1, named UCB-Tuned, which incorporates estimates of the reward variances, and show it to outperform UCB1 in simulations, although no theoretical results are given for the variant.

Two recently-proposed variants of the UCB1 algorithm are the MOSS (Minimax Optimal Strategy in the Stochastic case) algorithm (Audibert and Bubeck, 2010) and the UCB-V algorithm (Audibert and Bubeck, 2009). The MOSS algorithm is defined for finite problems with known horizon $|\mathcal{T}|$, but the 'doubling trick' described in §2.3 of Cesa-Bianchi and Lugosi (2006) can be used if the horizon is not known. MOSS differs from UCB1 by replacing the $\log(t-1)$ term in the exploratory value with $\log\left(\frac{|\mathcal{T}|}{|\mathcal{A}|n_{t,a}}\right)$ and hence selecting intensively drawn actions less often. The UCB-V algorithm incorporates estimates of reward variance in a similar way to the UCB-Tuned algorithm. The UCB-Tuned, MOSS and UCB-V algorithms provide suitable benchmarks for comparison in Bernoulli bandit problems.

Another class of 'UCB-type' algorithms was proposed initially by Lai and Robbins (1985), with a recent theoretical analysis by Garivier and Cappé (2011). The evaluation of action values involves constrained maximisation of Kullback-Leibler divergences. The primary purpose of the KL-UCB algorithm is to address the non-parametric problem although parametric implementation is discussed and optimal asymptotic regret bounds are proven for Bernoulli rewards. In the parametric case, a total action value corresponds to the highest posterior mean associated with a posterior distribution that has KL divergence less than a pre-defined term increasing logarithmically with time. A variant of KL-UCB, named KL-UCB+ is also proposed by Garivier and Cappé (2011) and is shown to outperform KL-UCB (with respect to expected regret) in simulated Bernoulli reward problems. Both algorithms also serve as suitable benchmarks for comparison in Bernoulli bandit problems.

For contextual bandit problems, Interval estimation (IE) methods, such as those suggested by Kaelbling (1994), Pavlidis et al. (2008) and Li et al. (2010) (under the name LinUCB), have become popular. They are UCB-type methods in which actions are selected greedily based on the upper bound of a confidence interval for the exploitative value estimate at a fixed significance level. The exploratory value used in IE methods is the difference between the upper bound and the exploitative value estimate. The width of the confidence interval at a particular point in the regressor space is expected to decrease the more times the action is selected.

There are numerous finite-time analyses of the contextual bandit problem. The case of linear expected reward functions provides the simplest contextual setting and examples of finite-time analyses include those of the SupLinRel and SupLinUCB algorithms by Auer (2002) and Chu et al. (2011) respectively, in which high probability regret bounds are established. The case of generalised linear expected rewards is considered by Filippi et al. (2010), proving high probability regret bounds for the GLM-UCB algorithm. Slivkins (2011) provides an example of finite-time analysis of contextual bandits in a more general setting, in which a regret bound is proved for the Contextual Zooming algorithm under the assumptions that the joint regressor and action space is a compact metric space and the reward functions are Lipschitz continuous over the aforementioned space.

On the other hand, very little is known about the theoretical properties of Thompson sampling. The only theoretical studies of Thompson sampling that we are aware of are by Granmo (2008)

and Agrawal and Goyal (2011). The former work considers only the two-armed non-contextual Bernoulli bandit and proves that Thompson sampling (the Bayesian Learning Automaton, in their terminology) converges to only pulling the optimal action with probability one. The latter work considers the K-armed non-contextual Bernoulli bandit and proves an optimal rate of regret (uniformly through time) for Thompson sampling. In this work, we focus on proving convergence criterion (2) for the LTS and OBS algorithms in a general contextual bandit setting in §3 and perform numerical experiments in §4 to illustrate the finite time properties of the algorithms.

## 2. Algorithms

In this section, we describe explicitly how the action selection is carried out at each decision instant for both the LTS and the OBS algorithms.

At each time $t$, the LTS algorithm requires a mechanism that can, for each action $a \in \mathcal{A}$, be used to sample from the posterior distribution of $f_a(x_t)$ given regressor $x_t$ and information set $I_t$. Recall that the density of this distribution is denoted as $p_a(\cdot|\, I_t, x_t)$ and a random variable from the distribution as $Q_{t,a}^{\mathrm{Th}}$.

**Algorithm 1** Local Thompson Sampling (LTS)

**Input:** Posterior distributions $\{p_a(\cdot|\, I_t, x_t) : a \in \mathcal{A}\}$
**for** $a = 1$ **to** $A$ **do**
  Sample $Q_{t,a}^{\mathrm{Th}} \sim p_a(\cdot|\, I_t, x_t)$
**end for**
Sample $a_t$ uniformly from $\operatorname{argmax}_{a\in\mathcal{A}} Q_{t,a}^{\mathrm{Th}}$

As in the case of the LTS algorithm, at each time $t$, the OBS algorithm requires a mechanism that can, for each action $a \in \mathcal{A}$, be used to sample from the posterior distribution of $f_a(x_t)$ given regressor $x_t$ and information set $I_t$. Additionally, the OBS algorithm requires a mechanism for evaluating exploitative value $\hat{f}_{t,a}(x_t)$, where exploitative value is taken to be the posterior expectation of $f_a(x_t)$ given $I_t$ and $x_t$.

**Algorithm 2** Optimistic Bayesian Sampling (OBS)

**Input:** Posterior distributions $\{p_a(\cdot|\, I_t, x_t) : a \in \mathcal{A}\}$
**for** $a = 1$ **to** $A$ **do**
  Sample $Q_{t,a}^{\mathrm{Th}} \sim p_a(\cdot|\, I_t, x_t)$
  Evaluate $\hat{f}_{t,a}(x_t) = \mathbb{E}(Q_{t,a}^{\mathrm{Th}}|\, I_t, x_t)$
  Set $Q_{t,a} = \max(Q_{t,a}^{\mathrm{Th}}, \hat{f}_{t,a}(x_t))$
**end for**
Sample $a_t$ uniformly from $\operatorname{argmax}_{a\in\mathcal{A}} Q_{t,a}$

## 3. Analysis

Theoretical properties of the LTS and OBS algorithms are analysed in this section. In particular, we focus on proving convergence in the sense of (2) under mild assumptions on the posterior distributions and expectations used. Regret analysis would provide useful insight into the finite time

properties of the LTS and OBS algorithms. However, we consider the problem in a general setting and impose only weak constraints on the nature of the posterior distributions used to sample action values, making the type of regret analysis common for UCB methods difficult, but allowing the convergence result to hold for a wide class of bandit settings and posterior approximations.

### 3.1 LTS Algorithm Analysis

We begin our convergence analysis by showing that the LTS algorithm explores all actions infinitely often, thus allowing a regression procedure to estimate all the functions $f_a$. In order to do this we need to make some assumptions.

To guarantee infinite exploration, it is desirable that the posterior distributions, $p_a(\cdot|\cdot,\cdot)$, generating the LTS samples are supported on $(\inf\mathcal{S},\sup\mathcal{S})$, a reasonable assumption in many cases. We make the weaker assumption that each sample can be greater than (or less than) any value in $(\inf\mathcal{S},\sup\mathcal{S})$ with positive probability. For instance, this assumption is satisfied by any distribution supported on $(\inf\mathcal{S},\inf\mathcal{S}+\delta_1)\cup(\sup\mathcal{S}-\delta_2,\sup\mathcal{S})$ for $\delta_1,\delta_2>0$.

It is also desirable that the posterior distributions remain fixed in periods of time in which the associated action is not selected, also a reasonable assumption if inference is independent for different actions. We make the weaker assumption that, in such periods of time, a lower bound exists for the probability that the LTS sample is above (or below) any value in $(\inf\mathcal{S},\sup\mathcal{S})$. Formally, we make the following assumption:

**Assumption 1** *Let $a\in\mathcal{A}$ be an arbitrary action, let $T$ be an arbitrary time, let $I_T$ be an arbitrary history to time $T$, and let $M\in(\inf\mathcal{S},\sup\mathcal{S})$. There exists an $\varepsilon>0$ depending on $a$, $T$, $I_T$ and $M$ such that for all $t>T$, all histories*

$$I_t = I_T\cup\{x_T,\ldots,x_{t-1},r_T,\ldots,r_{t-1},a_T,\ldots,a_{t-1}\}$$

*such that $a_s\neq a$ for $s\in\{T,\ldots,t-1\}$, and all $x_t\in\mathcal{X}$*

$$\mathbb{P}(Q^{Th}_{t,a}>M|I_t,x_t)>\varepsilon$$

*and*

$$\mathbb{P}(Q^{Th}_{t,a}<M|I_t,x_t)>\varepsilon.$$

Along with Assumption 1, we also assume that the posterior distributions concentrate on functions of the regressor bounded away from $\sup\mathcal{S}$ as their associated actions are selected infinitely often. Formally, we assume that:

**Assumption 2** *For each action $a\in\mathcal{A}$, there exist a function $g_a:\mathcal{X}\to(\inf\mathcal{S},\sup\mathcal{S})$ such that*

*(i)* $\left[Q^{Th}_{t,a}-g_a(x_t)\right]\xrightarrow{\mathbb{P}}0$ *as* $n_{t,a}\to\infty$,

*(ii)* $\sup_{x\in\mathcal{X}}g_a(x)<\sup\mathcal{S}$.

We do not take $g_a=f_a$ since this allows us to prove infinite exploration even when our regression framework does not support the true functions (e.g., when $I_0$ supports only linear functions, but the true $f_a$ are actually non-linear functions). Furthermore, the second condition, when combined with Assumption 1, ensures that over periods in which action $a$ is not selected there is a constant lower

bound on the probability that either the LTS or OBS algorithms sample a $Q_{t,a}^{\text{Th}}$ value greater than any $g_{\tilde{a}}(x)$.

Although there is an apparent tension between Assumption 1 and Assumption 2(i), note that Assumption 1 applies to the support of the posterior distributions for periods in which associated actions are not selected, whereas Assumption 2(i) applies to the limits of the posterior distributions as their associated actions are selected infinitely often.

Lemma 2 shows that, if Assumption 1 and 2 hold, then the proposed algorithm does guarantee infinite exploration. The lemma is important as it can be combined with Assumption 2 to imply that, for all $a \in \mathcal{A}$,

$$\left[Q_{t,a}^{\text{Th}} - g_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty$$

since $\forall a \in \mathcal{A}, n_{t,a} \to \infty$ as $t \to \infty$. The proof of Lemma 2 relies on the following lemma (Corollary 5.29 of Breiman, 1992):

**Lemma 1** *(Extended Borel-Cantelli Lemma). Let $I_t$ be an increasing sequence of $\sigma$-fields and let $V_t$ be $I_{t+1}$-measurable. Then*

$$\left\{\omega : \sum_{t=0}^{\infty} \mathbb{P}(V_t | I_t) = \infty\right\} = \{\omega : \omega \in V_t \textit{ infinitely often}\}$$

*holds with probability* 1.

**Lemma 2** *If Assumption 1 and 2 hold, then the LTS algorithm exhibits infinite exploration with probability* 1*, that is,*

$$\mathbb{P}\left(\bigcup_{a \in \mathcal{A}} \{n_{t,a} \to \infty \textit{ as } t \to \infty\}\right) = 1.$$

**Proof** Fix some arbitrary $k \in \{2, \ldots, A\}$. Assume without loss of generality that actions in $\mathcal{A}^{\text{inf}} = \{k, \ldots, A\}$ are selected infinitely often and actions in $\mathcal{A}^{\text{fin}} = \{1, \ldots, k-1\}$ are selected finitely often. By Assumption 2 and the infinite exploration of actions in $\mathcal{A}^{\text{inf}}$, we have that for all actions $a^{\text{inf}} \in \mathcal{A}^{\text{inf}}$ there exists a function $g_{a^{\text{inf}}} : \mathcal{X} \to (\inf \mathcal{S}, \sup \mathcal{S})$ such that

$$\left[Q_{t,a^{\text{inf}}}^{\text{Th}} - g_{a^{\text{inf}}}(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty.$$

Therefore, for fixed $\delta > 0$, there exists a finite random time, $T_\delta$, that is the earliest time in $\mathcal{T}$ such that for all actions $a^{\text{inf}} \in \mathcal{A}^{\text{inf}}$ we have

$$\mathbb{P}\left(|Q_{t,a^{\text{inf}}}^{\text{Th}} - g_{a^{\text{inf}}}(x_t)| < \delta \middle| I_t, x_t, t > T_\delta\right) > 1 - \delta. \tag{3}$$

Note that, by Assumption 2, we can choose $\delta$ to be small enough that such that for all actions $a \in \mathcal{A}$ and regressors $x \in \mathcal{X}$,

$$g_a(x) + \delta < \sup \mathcal{S}. \tag{4}$$

Since all actions in $\mathcal{A}^{\text{fin}}$ are selected finitely often, there exists some finite random time $T_f$ that is the earliest time in $\mathcal{T}$ such that no action in $\mathcal{A}^{\text{fin}}$ is selected after $T_f$. Let $T = \max\{T_\delta, T_f\}$. From (4) and Assumption 1 we have that for each $a^{\text{fin}} \in \mathcal{A}^{\text{fin}} \backslash 1$ there exists an $\varepsilon_{a^{\text{fin}}} > 0$ such that

$$\mathbb{P}\left(Q_{t,a^{\text{fin}}}^{\text{Th}} < \max_{a \in \mathcal{A}} g_a(x_t) + \delta \middle| I_t, x_t, t > T\right) > \varepsilon_{a^{\text{fin}}}, \tag{5}$$

and also that there exists an $\varepsilon_1 > 0$ such that

$$\mathbb{P}\left( Q_{t,1}^{\mathrm{Th}} > \max_{a \in \mathcal{A}} g_a(x_t) + \delta \middle| I_t, x_t, t > T \right) > \varepsilon_1. \tag{6}$$

Define the events:

$$\begin{aligned}
\overline{G}_{t,a}^{\delta} &= \left\{ Q_{t,a}^{\mathrm{Th}} > \max_{a \in \mathcal{A}} g_a(x_t)) + \delta \right\}, \\
\underline{G}_{t,a}^{\delta} &= \left\{ Q_{t,a}^{\mathrm{Th}} < \max_{a \in \mathcal{A}} g_a(x_t) + \delta \right\}, \\
G_{t,a}^{\delta} &= \left\{ |Q_{t,a}^{\mathrm{Th}} - g_a(x_t)| < \delta \right\}.
\end{aligned}$$

Then the LTS action selection rule implies that

$$\overline{G}_{t,1}^{\delta} \cap \left( \bigcap_{a^{\mathrm{fin}} \in \mathcal{A}^{\mathrm{fin}} \setminus 1} \underline{G}_{t,a^{\mathrm{fin}}}^{\delta} \right) \cap \left( \bigcap_{a^{\mathrm{inf}} \in \mathcal{A}^{\mathrm{inf}}} G_{t,a^{\mathrm{inf}}}^{\delta} \right) \subset \{a_t = a\},$$

so that

$$\mathbb{P}(a_t = 1 | I_t, x_t, T > t) \geq \mathbb{P}\left( \overline{G}_{t,1}^{\delta} \cap \left( \bigcap_{a^{\mathrm{fin}} \in \mathcal{A}^{\mathrm{fin}} \setminus 1} \underline{G}_{t,a^{\mathrm{fin}}}^{\delta} \right) \cap \left( \bigcap_{a^{\mathrm{inf}} \in \mathcal{A}^{\mathrm{inf}}} G_{t,a^{\mathrm{inf}}}^{\delta} \right) \middle| I_t, x_t, t > T \right). \tag{7}$$

The set $\{\overline{G}_{t,1}^{\delta}, \underline{G}_{t,a}^{\delta}, G_{t,b}^{\delta} : a = 2, \ldots, k-1, b = k, \ldots, A\}$ is a conditionally independent set of events given $I_t$ and $x_t$. Therefore, by (3), (5) and (6), we have

$$\mathbb{P}\left( \overline{G}_{t,1}^{\delta} \cap \left( \bigcap_{a^{\mathrm{fin}} \in \mathcal{A}^{\mathrm{fin}} \setminus 1} \underline{G}_{t,a^{\mathrm{fin}}}^{\delta} \right) \cap \left( \bigcap_{a^{\mathrm{inf}} \in \mathcal{A}^{\mathrm{inf}}} G_{t,a^{\mathrm{inf}}}^{\delta} \right) \middle| I_t, x_t, t > T \right) > \varepsilon^{k-1}(1-\delta)^{A-k+1} \tag{8}$$

where $\varepsilon = \min_{a^{\mathrm{fin}} \in \mathcal{A}^{\mathrm{fin}}} \varepsilon_{a^{\mathrm{fin}}}$. Combining (7) and (8), it follows that

$$\mathbb{P}(a_t = 1 | I_t, x_t, t > T) > \varepsilon^{k-1}(1-\delta)^{A-k+1}$$

so that

$$\begin{aligned}
\sum_{t \in \mathcal{T}} \mathbb{P}(a_t = 1 | I_t, x_t) &\geq \sum_{t=T+1}^{\infty} \mathbb{P}(a_t = 1 | I_t, x_t) \\
&= \sum_{t=T+1}^{\infty} \mathbb{P}(a_t = 1 | I_t, x_t, t > T) \\
&> \sum_{t=T+1}^{\infty} \varepsilon^{k-1}(1-\delta)^{A-k+1} = \infty
\end{aligned}$$

since $T$ is almost surely finite. Hence, by Lemma 1, $\{a_t = 1\}$ occurs infinitely often almost surely, contradicting the assumption that $1 \in \mathcal{A}^{\mathrm{fin}}$. Since action 1 was chosen arbitrarily from the set $\mathcal{A}^{\mathrm{fin}}$,

any action in $\mathcal{A}^{\text{fin}}$ would cause a contradiction. Therefore, $\mathcal{A}^{\text{fin}} = \emptyset$, that is, every action is selected infinitely often almost surely. ■

When we return to the notion of exploitative value estimates $\hat{f}_{t,a}(x_t)$ and hence the concept of a greedy action, then we wish to ascertain whether the algorithm is greedy in the limit. Assumption 2 only implies that the sum of exploitative and exploratory values tends to a particular function of the regressor and not that the exploratory values tend to zero. Although a minor point, the infinite exploration, given by Assumption 1 and 2, needs to be complemented with an assumption that the exploitative value estimates are converging to the same limit as the sampled values $Q_{t,a}^{\text{Th}}$ in order to prove that the policy generated by the LTS algorithm is GLIE. This assumption is not used in proving that the LTS algorithm generates policies satisfying convergence criterion (2) but is used for the equivalent proof for the OBS algorithm (see §3.2).

**Assumption 3** *For all actions $a \in \mathcal{A}$*

$$\left[\hat{f}_{t,a}(x_t) - g_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } n_{t,a} \to \infty$$

*for $g_a$ defined as in Assumption 2.*

**Lemma 3** *If Assumptions 1, 2 and 3 hold, then the LTS algorithm policy is GLIE.*

**Proof** For any $a \in \mathcal{A}$, since

$$\tilde{f}_{t,a}^{\text{Th}} = Q_{t,a}^{\text{Th}} - \hat{f}_{t,a}(x_t),$$

Assumptions 2 and 3 give

$$\tilde{f}_{t,a}^{\text{Th}}(x_t) \xrightarrow{\mathbb{P}} 0 \text{ as } n_{t,a} \to \infty. \tag{9}$$

Since Assumptions 1 and 2 are satisfied, infinite exploration is guaranteed by Lemma 2. This infinite exploration and (9) imply that $\forall a \in \mathcal{A}$

$$\tilde{f}_{t,a}^{\text{Th}}(x_t) \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty. \tag{10}$$

Let us denote the set

$$\mathcal{A}_t^* = \operatorname*{argmax}_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t).$$

By splitting value samples into exploitative and exploratory components we have

$$\begin{aligned}
\mathbb{P}\left(a_t \in \mathcal{A}_t^* \middle| I_t, x_t\right) &= \mathbb{P}\left(\max_{a \in \mathcal{A}_t^*} Q_{t,a}^{\text{Th}} > \max_{a \in \mathcal{A} \setminus \mathcal{A}_t^*} Q_{t,a}^{\text{Th}} \middle| I_t, x_t\right) \\
&= \mathbb{P}\left(\max_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t) + \max_{a \in \mathcal{A}_t^*} \tilde{f}_{t,a}^{\text{Th}}(x_t) > \max_{a \in \mathcal{A} \setminus \mathcal{A}_t^*} \left[\hat{f}_{t,a}(x_t) + \tilde{f}_{t,a}^{\text{Th}}(x_t)\right] \middle| I_t, x_t\right) \\
&\geq \mathbb{P}\left(\max_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t) - \max_{a \in \mathcal{A} \setminus \mathcal{A}_t^*} \hat{f}_{t,a}(x_t) > 2 \max_{a \in \mathcal{A}} \left|\tilde{f}_{t,a}^{\text{Th}}(x_t)\right| \middle| I_t, x_t\right) \\
&\xrightarrow{\text{a.s.}} 1 \text{ as } t \to \infty,
\end{aligned}$$

since the right hand side of the last inequality converges in probability to 0 by (10) and

$$\max_{a \in \mathcal{A}} \hat{f}_{t,a}(x_t) > \max_{a \in \mathcal{A} \setminus \mathcal{A}_t^*} \hat{f}_{t,a}(x_t)$$

by definition of $\mathcal{A}_t^*$. Hence, the action selection is greedy in the limit. Lemma 2 ensures infinite exploration, so the policy is GLIE. ■

We have shown we can achieve a GLIE policy even when we do not have consistent regression. However, to ensure the convergence condition (2) is satisfied we need to assume consistency, that is, that the functions $g_a$ (to which the $Q_{t,a}^{\text{Th}}$ converge) are actually the true functions $f_a$.

**Assumption 4** *For all actions $a \in \mathcal{A}$ and regressors $x \in \mathcal{X}$,*

$$g_a(x) = f_a(x).$$

The following Theorem is the main convergence result for the LTS algorithm. Its proof uses the fact that, under the specified assumptions, Lemma 2 implies that, for all actions $a \in \mathcal{A}$,

$$\left[Q_{t,a}^{\text{Th}} - f_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty.$$

We then use a coupling argument (dealing with the dependence in the action selection sequence) to prove that the LTS algorithm policy satisfies convergence criterion (2).

**Theorem 1** *If Assumptions 1, 2 and 4 hold, then the LTS algorithm will produce a policy satisfying convergence criterion (2).*

**Proof** Recall that the optimal expected reward function is defined by $f^*(x) = \max_{a \in \mathcal{A}} f_a(x)$. Fix some arbitrary $\delta > 0$. Denote the event

$$E_t^\delta = \left\{ f^*(x_t) - f_{a_t}(x_t) < 2\delta \right\}$$

so that $E_t^\delta$ is the event that true expected reward for the action selected at time $t$ is within $2\delta$ of the optimal expected reward at time $t$.

The first part of the proof consists of showing that

$$\mathbb{P}(E_t^\delta | I_t, x_t) \xrightarrow{\text{a.s.}} 1 \text{ as } t \to \infty.$$

From Assumptions 2 and 4, and the infinite exploration guaranteed by Lemma 2, $\forall a \in \mathcal{A}$

$$\left[Q_{t,a}^{\text{Th}} - f_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty.$$

Therefore there exists a finite random time, $T_\delta$, that is the earliest time in $\mathcal{T}$ such that $\forall a \in \mathcal{A}$

$$\mathbb{P}\left( |Q_{t,a}^{\text{Th}} - f_a(x_t)| < \delta \middle| I_t, x_t, t > T_\delta \right) > 1 - \delta \tag{11}$$

so that, after $T_\delta$, all sampled $Q_{t,a}^{\text{Th}}$ values are within $\delta$ of the true values with high probability.

Define the events

$$F_{t,a}^\delta = \{ |Q_{t,a}^{\text{Th}} - f_a(x_t)| < \delta \}.$$

Then $\{F_{t,a}^{\delta} : a \in \mathcal{A}\}$ is a conditionally independent set of events given $I_t$ and $x_t$, so that

$$\mathbb{P}\left(\bigcap_{a\in\mathcal{A}} F_{t,a}^{\delta} | I_t, x_t, t > T_{\delta}\right) = \prod_{a\in\mathcal{A}} \mathbb{P}(F_{t,a}^{\delta} | I_t, x_t, t > T_{\delta}) > (1-\delta)^A \tag{12}$$

using inequality (11).

Note that, for any $a_t^* \in \operatorname{argmax}_{a\in\mathcal{A}} f_a(x_t)$,

$$\bigcap_{a\in\mathcal{A}} F_{t,a}^{\delta} \subset \{Q_{t,a_t^*}^{\mathrm{Th}} > f^*(x_t) - \delta\} \tag{13}$$

and, for any $a' \in \{a \in \mathcal{A} : f^*(x_t) - f_a(x_t) > 2\delta\}$,

$$\bigcap_{a\in\mathcal{A}} F_{t,a}^{\delta} \subset \{Q_{t,a'}^{\mathrm{Th}} < f^*(x_t) - \delta\}. \tag{14}$$

Since $\operatorname{argmax}_{a\in\mathcal{A}} f_a(x_t)$ is non-empty and the action selection rule is greedy on the $Q_{t,a}^{\mathrm{Th}}$, statements (13) and (14) give

$$\bigcap_{a\in\mathcal{A}} F_{t,a}^{\delta} \subset \{f^*(x_t) - f_{a_t}(x_t) < 2\delta\} = E_t^{\delta}.$$

and so

$$\mathbb{P}\left(\bigcap_{a\in\mathcal{A}} F_{t,a}^{\delta} | I_t, x_t\right) \leq \mathbb{P}(E_t^{\delta} | I_t, x_t). \tag{15}$$

Inequalities (12) and (15) imply that

$$\mathbb{P}(E_t^{\delta} | I_t, x_t, t > T_{\delta}) > (1-\delta)^A.$$

The condition above holds for arbitrarily small $\delta$ so that $\forall x \in \mathcal{X}$

$$\mathbb{P}(E_t^{\delta} | I_t, x_t) \xrightarrow{\text{a.s.}} 1 \text{ as } t \to \infty. \tag{16}$$

This concludes the first part of the proof. We have shown that the probability that the action selected at time $t$ has a true expected reward that is within $2\delta$ of that of the action with the highest true expected reward at time $t$ tends to 1 as $t \to \infty$. We now face the difficulty that the strong law of large numbers cannot be used directly to establish a lower bound on $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} f_{a_s}(x_s)$ since the expected reward sequence $\left(f_{a_s}(x_s)\right)_{s\in\mathcal{T}}$ is a sequence of dependent random variables.

The result may be proved using a coupling argument. We will construct an independent sequence of actions $b_s$ that are coupled with $a_s$, but for which we can apply the strong law of large numbers to $f_{b_s}(x_s)$. By relating the expected reward for playing the $b_s$ sequence to that of the $a_s$ sequence we will show that the $a_s$ sequence satisfies the optimality condition (2).

Fix some arbitrary $\varepsilon > 0$, define the sets

$$\mathcal{A}_t^{\varepsilon} = \{a \in \mathcal{A} : f^*(x_t) - f_a(x_t) < 2\varepsilon\},$$

and let $U_1, U_2, \ldots$ be a sequence of independent and identically distributed $U[0,1]$ random variables. The construction of $E_s^{\varepsilon}$ and $\mathcal{A}_s^{\varepsilon}$ implies that $E_s^{\varepsilon} \Leftrightarrow \{a_s \in \mathcal{A}_s^{\varepsilon}\}$. So, by conditioning on the event $\{a_s \in \mathcal{A}_s^{\varepsilon}\}$ and using the LTS action selection rule, it follows that $a_s$ can be expressed as

$$a_s = \begin{cases} \operatorname{argmax}_{a\in\mathcal{A}_s^{\varepsilon}} Q_{s,a}^{\mathrm{Th}} & \text{if } U_s < \mathbb{P}(E_s^{\varepsilon} | I_s, x_s) \\ \operatorname{argmax}_{a\in\mathcal{A}\setminus\mathcal{A}_s^{\varepsilon}} Q_{s,a}^{\mathrm{Th}} & \text{if } U_s > \mathbb{P}(E_s^{\varepsilon} | I_s, x_s) \end{cases}$$

with ties resolved using uniform sampling.

We similarly define $b_s$ based on the $U_s$ as

$$b_s = \begin{cases} \operatorname{argmin}_{a \in \mathcal{A}_s^\varepsilon} f_a(x_s) & \text{if } U_s < 1-\varepsilon \\ \operatorname{argmin}_{a \in \mathcal{A}} f_a(x_s) & \text{if } U_s > 1-\varepsilon, \end{cases}$$

again, with ties resolved using uniform sampling. Note that, since the $U_s$ and $x_s$ are independent and identically distributed, the $b_s$ are independent and identically distributed, and so is the sequence $f_{b_s}(x_s)$.

Note that by (16) there exists a finite random time

$$S_\varepsilon = \sup\left\{ t < \infty : \mathbb{P}(E_t^\varepsilon | I_t, x_t) < 1-\varepsilon \right\}.$$

By considering the definition of $S_\varepsilon$, it follows that

$$\begin{aligned} \{s > S_\varepsilon\} \cap \{U_s < 1-\varepsilon\} &\subset \{U_s < \mathbb{P}(E_s^\varepsilon | I_s, x_s)\} \\ &\subset \left\{ a_s \in \operatorname*{argmax}_{a \in \mathcal{A}_s^\varepsilon} Q_{s,a}^{\mathrm{Th}} \right\} \\ &\subset \{a_s \in \mathcal{A}_s^\varepsilon\} \\ &\subset \left\{ f_{a_s}(x_s) \geq \min_{b \in \mathcal{A}_s^\varepsilon} f_b(x_s) \right\}. \end{aligned} \tag{17}$$

Also, it is the case that

$$\begin{aligned} \{U_s < 1-\varepsilon\} &= \left\{ b_s \in \operatorname*{argmin}_{b \in \mathcal{A}_s^\varepsilon} f_b(x_s) \right\} \\ &\subset \left\{ f_{b_s}(x_s) = \min_{b \in \mathcal{A}_s^\varepsilon} f_b(x_s) \right\}. \end{aligned} \tag{18}$$

Combining (17) and (18), we have that

$$\{s > S_\varepsilon\} \cap \{U_s < 1-\varepsilon\} \subset \left\{ f_{a_s}(x_s) \geq f_{b_s}(x_s) \right\}. \tag{19}$$

Note also that

$$\{U_s > 1-\varepsilon\} \subset \left\{ f_{b_s}(x_s) = \min_{a' \in \mathcal{A}} f_{a'}(x_s) \leq f_{a_s}(x_s) \right\}. \tag{20}$$

It follows from (19), (20) and the definition of $f^*$ that

$$\{s > S_\varepsilon\} \subset \{f^*(x_s) \geq f_{a_s}(x_s) \geq f_{b_s}(x_s)\}$$

and so

$$\frac{1}{t} \sum_{s=S_\varepsilon}^{t} f^*(x_s) \geq \frac{1}{t} \sum_{s=S_\varepsilon}^{t} f_{a_s}(x_s) \geq \frac{1}{t} \sum_{s=S_\varepsilon}^{t} f_{b_s}(x_s). \tag{21}$$

We will now use inequality (21) to prove the result. The definition of $b_s$ implies that

$$\{U_s < 1-\varepsilon\} \subset \{b_s \in \mathcal{A}_s^\varepsilon\}.$$

By considering the definition of $\mathcal{A}_s^\varepsilon$, it follows that

$$\{U_s < 1-\varepsilon\} \subset \{f_{b_s}(x_s) > f^*(x_s) - 2\varepsilon\}. \tag{22}$$

Since

- $S_\varepsilon$ is finite
- the $U_s$ are independent and identically distributed
- the $f_{b_s}(x_s)$ are independent and identically distributed

we can use the strong law of large numbers and (22) to get

$$\begin{aligned}\lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{b_s}(x_s)\right] &= \mathbb{E}_{\mathcal{U}\times\mathcal{X}} f_{b_s}(x_s)\\ &= \mathbb{P}(U_s < 1-\varepsilon)\mathbb{E}_{\mathcal{X}}[f_{b_s}(x_s)|U_s < 1-\varepsilon] + \mathbb{P}(U_s > 1-\varepsilon)\mathbb{E}_{\mathcal{X}}[f_{b_s}(x_s)|U_s > 1-\varepsilon]\\ &> (1-\varepsilon)\Big(\mathbb{E}_{\mathcal{X}} f^*(\cdot) - 2\varepsilon\Big) + \varepsilon\mathbb{E}_{\mathcal{X}}[f_{b_s}(x_s)|U_s > 1-\varepsilon], \end{aligned} \tag{23}$$

where $\mathbb{E}_{\mathcal{U}\times\mathcal{X}}$ denotes expectation taken with respect to the joint distribution of $U_t$ and $x_t$ and $\mathbb{E}_{\mathcal{X}}$ denotes expectation taken with respect to the distribution of $x_t$ (note that both distributions are the same for all values of $t$).

By the strong law of large numbers, we get

$$\lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f^*(x_s)\right] = \mathbb{E}_{\mathcal{X}} f^*(\cdot). \tag{24}$$

Since (21), (23) and (24) hold, we have that

$$\begin{aligned}\mathbb{E}_{\mathcal{X}} f^*(\cdot) &\geq \lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{a_s}(x_s)\right]\\ &\geq \lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{b_s}(x_s)\right]\\ &> (1-\varepsilon)\Big(\mathbb{E}_{\mathcal{X}} f^*(\cdot) - 2\varepsilon\Big) + \varepsilon\mathbb{E}_{\mathcal{X}}[f_{b_s}(x_s)|U_s > 1-\varepsilon].\end{aligned}$$

This holds for arbitrarily small $\varepsilon$, hence

$$\lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{a_s}(x_s)\right] = \mathbb{E}_{\mathcal{X}} f^*(\cdot). \tag{25}$$

It is the case that

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t} f_{a_s}(x_s) &= \lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{S_\varepsilon-1} f_{a_s}(x_s) + \lim_{t\to\infty}\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{a_s}(x_s)\\ &= 0 + \lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=S_\varepsilon}^{t} f_{a_s}(x_s)\right]\end{aligned} \tag{26}$$

as $t\to\infty$ since $S_\varepsilon$ is finite and $f_{a_s}(x_s) \leq \sup_{x\in\mathcal{X}} f_{a^*}(x) < \infty$.

Since both (25) and (26) hold, it is true that

$$\lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=1}^{t} f_{a_s}(x_s)\right] = \mathbb{E}_{\mathcal{X}} f^*(\cdot) = \lim_{t\to\infty}\left[\frac{1}{t}\sum_{s=1}^{t} f^*(x_s)\right].$$

Hence

$$\frac{\sum_{s=1}^{t} f_{a_s}(x_s)}{\sum_{s=1}^{t} f^*(x_s)} \xrightarrow{\text{a.s.}} 1 \text{ as } t \to \infty.$$

■

### 3.2 OBS Algorithm Analysis

We analyse the OBS algorithm in a similar way to the LTS algorithm. In order to prove infinite exploration for the OBS algorithm, we must make an additional assumption on the exploitative value estimates. We assume that exploitative values are less than $\sup \mathcal{S}$ by a constant for all regressor values during periods of time in which their associated actions are not selected. This allows us to make statements similar to inequality (5) in the proof of Lemma 2, however relating to OBS samples rather than LTS samples.

**Assumption 5** *Let $a \in \mathcal{A}$ be an arbitrary action, let $T$ be an arbitrary time, and let $I_T$ be an arbitrary history to time $T$. There exists a $\delta > 0$ depending on $a$, $T$, and $I_T$ such that for all $t > T$, all histories $I_t = I_T \cup \{x_T, \ldots, x_{t-1}, r_T, \ldots, r_{t-1}, a_T, \ldots, a_{t-1}\}$ such that $a_s \neq a$ for $s \in \{T, \ldots, t-1\}$, and all $x \in \mathcal{X}$,*

$$\sup \mathcal{S} - \hat{f}_{t,a}(x) > \delta.$$

We now show that the OBS algorithm explores all actions infinitely often. Assumptions 2 and 3 imply that, for any action $a \in \mathcal{A}$,

$$\left[Q_{t,a} - g_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } n_{t,a} \to \infty$$

so that OBS samples associated with actions assumed to be selected infinitely often can be treated in the same way as LTS samples are in the proof of Lemma 2. The only slight difference in the proof comes in the treatment of samples associated with actions assumed to be selected finitely often, although Assumption 5 ensures that the logic is similar.

**Lemma 4** *If Assumption 1, 2, 3 and 5 hold, then the OBS algorithm exhibits infinite exploration with probability* 1.

**Proof** Since $Q_{t,a} = \max(Q^{\text{Th}}_{t,a}, \hat{f}_{t,a}(x_t))$, Assumption 2 and 3 give that $\forall a \in \mathcal{A}^{\text{inf}}$

$$\left[Q_{t,a} - g_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty.$$

Let $T$ and $\delta$ be defined as in Lemma 2 (with the $Q^{\text{Th}}_{t,a}$ replaced by $Q_{t,a}$). In the proof of Lemma 2, $g^*(x_t) := \max_{a \in \mathcal{A}} g_a(x_t) + \delta$ is used as a target for samples associated with actions in $a^{\text{fin}} \in \mathcal{A} \backslash 1$ to fall below and the sample associated with action 1 to fall above. The assumptions do not restrict from occurring the event that there exists an action $a$ in $\mathcal{A}^{\text{fin}} \backslash 1$ such that, for all $t > T$, $\hat{f}_{t,a}(x_t) > g^*(x_t)$, thus making it impossible for $Q_{t,a}$ to fall below $g^*(x_t)$. However, Assumption 5 can be used to imply that there exists a $\delta_1 > 0$ such that $\forall a^{\text{fin}} \in \mathcal{A}^{\text{fin}}$ and $\forall t > T$

$$\hat{f}_{t,a^{\text{fin}}}(x_t) < \sup \mathcal{S} - \delta_1. \tag{27}$$

Assumption 1 and inequality (27) then imply that, for all actions $a^{\text{fin}} \in \mathcal{A}^{\text{fin}} \setminus 1$, there exists an $\varepsilon_{a^{\text{fin}}} > 0$ such that

$$\mathbb{P}\Big(Q_{t,a^{\text{fin}}} < \max(g^*(x_t), \sup \mathcal{S} - \delta_1) \Big| I_t, x_t, t > T\Big) > \varepsilon_{a^{\text{fin}}}$$

and also that there exists an $\varepsilon_1 > 0$ such that

$$\mathbb{P}\Big(Q_{t,1} > \max(g^*(x_t), \sup \mathcal{S} - \delta_1) \Big| I_t, x_t, t > T\Big) > \varepsilon_1.$$

The proof then follows in a similar manner to that of Lemma 2, with the $Q^{\text{Th}}_{t,a}$ replaced by $Q_{t,a}$. ■

In the case of the LTS algorithm, it is not necessary for the generated policy to be GLIE for Theorem 1 to hold. Assumptions are only made on total action value estimates, that is, the sum of exploitative and exploratory value, and it is not necessary that the exploratory value converges to zero. Exploitative value estimates are not used explicitly for the LTS algorithm and Lemma 3 is included in this work for completeness. In the case of the OBS algorithm, it is important that Assumption 3 holds so that the policy is GLIE, since exploitative values are used explicitly. The total action value can be equal to the exploitative value estimate so it is important that the exploitative estimate converges to the same value as the LTS samples. Obviously, this would hold if the posterior expectation is used as we suggest, however our framework allows for the use any functions of the regressor satisfying Assumptions 3 and 5 when implementing the OBS algorithm and the convergence result will still hold.

**Lemma 5** *If Assumption 1, 2, 3 and 5 hold, then the OBS algorithm policy is GLIE.*

**Proof** The proof is similar to that of Lemma 3, replacing $\tilde{f}^{\text{Th}}_{t,a}$ with $\tilde{f}_{t,a}$, replacing $Q^{\text{Th}}_{t,a}$ with $Q_{t,a}$ and using the fact that

$$\tilde{f}_{t,a}(x_t) = \max(0, \tilde{f}^{\text{Th}}_{t,a}(x_t)).$$

■

Under Assumptions 1–5, we have that the LTS samples, $Q^{\text{Th}}_{t,a}$, and the exploitative values, $\hat{f}_{t,a}(x_t)$ are consistent estimators of the true expected rewards, $f_a(x_t)$ and that infinite exploration is guaranteed by Lemma 4. Therefore, we have that the OBS samples, $Q_{t,a}$ converge in probability to the true expected rewards, $f_a(x_t)$, as $t \to \infty$. We can therefore prove that the OBS algorithm satisfies convergence criterion (2) using a similar method to that used for the proof of Theorem 1.

**Theorem 2** *If Assumptions 1–5 hold, then the OBS algorithm will produce a policy satisfying convergence criterion (2).*

**Proof** By Assumption 2, 3 and 4 and the infinite exploration guaranteed by Lemma 4, we have that $\forall a \in \mathcal{A}$

$$\big[Q_{t,a} - f_a(x_t)\big] \xrightarrow{\mathbb{P}} 0 \text{ as } t \to \infty$$

since $Q_{t,a} = \max(Q^{\text{Th}}_{t,a}, \hat{f}_{t,a}(x_t))$. The remainder of the proof follows as in the case of Theorem 1 (replacing $Q^{\text{Th}}_{t,a}$ with $Q_{t,a}$). ■

## 4. Case Studies

In this section, we aim to validate claims made in §1.2 regarding the short term performance of the OBS algorithm by means of simulation. We use the notion of cumulative pseudo-regret (Filippi et al., 2010) to assess the performance of an algorithm. The cumulative pseudo-regret measures the expected difference between the reward the algorithm receives and the reward that would be received if the regression functions were known in advance so that an optimal arm can be chosen on every timestep; it is a standard measure of finite-time performance of a bandit algorithm. Our definition differs slightly from that of Filippi et al. (2010) since we do not restrict attention to generalised linear bandits.

**Definition 4** *The cumulative (pseudo) regret, $R_T$, at time $T$ is given by*

$$R_T = \sum_{t=1}^{T} \Big[ f^*(x_t) - f_{a_t}(x_t) \Big].$$

We compare the performance of OBS to that of LTS and various recently proposed action selection methods in simulated Bernoulli bandit and linear regression problem settings in §4.1 and §4.2 respectively. We also consider a real-world version of the problem using data that relates to personalised news article recommendation, the Yahoo! Front Page Today Module User Click Log Data Set (Yahoo! Academic Relations, 2011). Graepel et al. (2010) suggest using LTS to deal with the exploration-exploitation dilemma in a similar sponsored search advertising setting. We compare the OBS performance to that of LTS on the Yahoo! data and obtain results indicating that OBS performs better in the short term.

### 4.1 Bernoulli Bandit

In the multi-armed Bernoulli bandit problem, there is no regressor present. If the agent chooses action $a$ on any timestep then a reward of 1 is received with probability $p_a$ and 0 with probability $1-p_a$. For each action $a$, the probability $p_a$ can be estimated by considering the frequency of success observed in past selections of the action. The agent needs to explore in order to learn the probabilities of success for each action, so that the action yielding the highest expected reward can be identified. The agent needs to exploit what has been learned in order to maximise expected reward. The multi-armed Bernoulli bandit problem presents a simple example of the exploration-exploitation dilemma, and has therefore been studied extensively.

#### 4.1.1 PROBLEM CONSIDERED

In this case, we let the prior information, $I_0$, consist of the following:

- The number of actions, $A$.
- $(\forall a \in \mathcal{A})(\forall t \in \mathcal{T})\big\{ f_a(x_t) = p_a \big\}$ for $p_a \in (0,1)$ unknown.
- $\forall a, \forall t, z_{t,a} = \begin{cases} -p_a & \text{with probability } 1-p_a, \\ 1-p_a & \text{with probability } p_a. \end{cases}$
- For each action $a \in \mathcal{A}$, the prior distribution of $f_a$ is Beta$(1,1)$ (or equivalently U$(0,1)$).

#### 4.1.2 LTS AND OBS IMPLEMENTATION

Let $\tilde{r}_{\tau,a}$ denote the value of the reward received on the timestep where action $a$ was picked for the $\tau$th time. For arbitrary $a \in \mathcal{A}$ define

$$s_{t,a} = \sum_{\tau=1}^{n_{t,a}} \tilde{r}_{\tau,a}.$$

Posterior expectations (using flat priors, as indicated by $I_0$) can be evaluated easily, so we define exploitative value as

$$\hat{f}_{t,a} := \frac{s_{t,a}+1}{n_{t,a}+2}.$$

The posterior distribution of $p_a$ given $I_t$ has a simple form. We sample

$$Q^{\mathrm{Th}}_{t,a} \sim \mathrm{Beta}(s_{t,a}+1, n_{t,a}-s_{t,a}+1).$$

and set

$$Q_{t,a} = \max(Q^{\mathrm{Th}}_{t,a}, \hat{f}_{t,a}).$$

#### 4.1.3 CONVERGENCE

In this section, we check explicitly that Assumptions 1–5 are satisfied in this Bernoulli bandit setting, therefore proving that the LTS and OBS algorithms generate policies satisfying convergence criterion (2).

**Lemma 6** *The LTS total value estimate, $Q^{Th}_{t,a}$, satisfies Assumption 1, for all $a \in \mathcal{A}$.*

**Proof** Let $a \in \mathcal{A}$, $T > 0$, $I_T$ and $M \in (0,1)$ be arbitrary. For any $t > T$ and $I_t = I_T \cup \{r_T, \ldots, r_{t-1}, a_T, \ldots, a_{t-1}\}$ with $a_s \neq a$ for $s \in \{T, \ldots, t-1\}$, the posterior distribution of $f_a$ given $I_t$ will be the same as the posterior distribution of $f_a$ given $I_T$ (since no further information about $f_a$ is contained in $I_t$). Let

$$\varepsilon := \frac{1}{2} \min\left\{ \mathbb{P}(Q^{\mathrm{Th}}_{T,a} < M \mid I_T), \mathbb{P}(Q^{\mathrm{Th}}_{T,a} > M \mid I_T) \right\}.$$

We then have that

$$\mathbb{P}(Q^{\mathrm{Th}}_{t,a} > M | I_t) > \varepsilon$$

and

$$\mathbb{P}(Q^{\mathrm{Th}}_{t,a} < M | I_t) > \varepsilon.$$

■

**Lemma 7** *The LTS total value estimate, $Q^{Th}_{t,a}$, satisfies Assumptions 2–4, for all $a \in \mathcal{A}$.*

**Proof** Posterior expectations are given by

$$\begin{aligned} \hat{f}_{t,a} &:= \frac{s_{t,a}+1}{n_{t,a}+2} \\ &= \frac{1+\sum_{\tau=1}^{n_{t,a}} \tilde{r}_{\tau,a}}{n_{t,a}+2}. \end{aligned}$$

Using the strong law of large numbers, we then have

$$\lim_{n_{t,a}\to\infty} \hat{f}_{t,a} = \lim_{n_{t,a}\to\infty} \frac{\sum_{\tau=1}^{n_{t,a}} \tilde{r}_{\tau,a}}{n_{t,a}} = \mathbb{E}(r_t|a_t = a) = p_a = f_a. \tag{28}$$

Therefore, it is the case that

$$\mathbb{E}(Q_{t,a}^{\mathrm{Th}}|\, I_t) = \hat{f}_{t,a} \xrightarrow{\text{a.s.}} f_a \text{ as } n_{t,a} \to \infty. \tag{29}$$

By considering the variance of the LTS samples, we get

$$\begin{aligned}
\mathrm{Var}(Q_{t,a}^{\mathrm{Th}}|\, I_t) &= \frac{(s_{t,a}+1)(n_{t,a}-s_{t,a}+1)}{(n_{t,a}+2)^2(n_{t,a}+3)} \\
&< \frac{(n_{t,a}+2)^2}{(n_{t,a}+2)^2(n_{t,a}+3)} \\
&= \frac{1}{(n_{t,a}+3)} \xrightarrow{\text{a.s.}} 0 \text{ as } n_{t,a} \to \infty.
\end{aligned} \tag{30}$$

From (29) and (30), we then have $\forall a \in \mathcal{A}$

$$Q_{t,a}^{\mathrm{Th}} \xrightarrow{\mathbb{P}} f_a \text{ as } n_{t,a} \to \infty. \tag{31}$$

Note that since $f_a = p_a < 1$ for each $a \in \mathcal{A}$, and $|\mathcal{A}| < \infty$, convergence result (31) shows that Assumptions 2 and 4 hold and convergence results (31) and (28) combined show that Assumption 3 holds. ■

**Lemma 8** *The exploitative value estimate, $\hat{f}_{t,a}$, satisfies Assumption 5, for all $a \in \mathcal{A}$.*

**Proof** Let $a \in \mathcal{A}$, $T > 0$ and $I_T$ be arbitrary. For any $t > T$ and $I_t = I_T \cup \{r_T, \ldots, r_{t-1}, a_T, \ldots, a_{t-1}\}$ with $a_s \neq a$ for $s \in \{T, \ldots, t-1\}$,

$$n_{t,a} = n_{T,a} \quad \text{and} \quad s_{t,a} = s_{T,a}.$$

Therefore

$$\hat{f}_{t,a} = \frac{s_{T,a}+1}{n_{T,a}+2} \leq \frac{n_{T,a}+1}{n_{T,a}+2} < 1 - \frac{1}{n_{T,a}+2} = \sup S - \frac{1}{n_{T,a}+2},$$

so that the assumption is satisfied with $\delta = \frac{1}{n_{T,a}+2}$. ■

**Proposition 1** *Within the described Bernoulli bandit setting convergence criterion (2) is satisfied when the LTS or the OBS algorithm is used.*

**Proof** Assumptions 1–5 hold, so the proof follows directly from Theorems 1 and 2. ■

#### 4.1.4 EXPERIMENTAL RESULTS

We parameterise a Bernoulli problem of the described form with a vector of probabilities, $(p_1, \ldots, p_A)$, corresponding to the expected rewards for the actions in $\mathcal{A}$. We simulate the problem in four environments with parameters $(0.8, 0.9)$, $(0.8, 0.8, 0.8, 0.9)$, $(0.45, 0.55)$ and $(0.45, 0.45, 0.45, 0.55)$. It is well known that the variance of a Bernoulli random variable is maximised when the associated probability of success is 0.5. We choose to consider the four environments mentioned to provide 'low variance' and 'high variance' versions of the problem and to investigate the effect of increasing the number of actions.

For each problem environment, the process is run for 8000 independent trials. A time window of $\mathcal{T} = \{1, \ldots, 5000\}$ is considered on each trial. A trial consists of sampling the potential rewards $r_{t,a} \sim \text{Bernoulli}(p_a)$ for each $t \in \mathcal{T}$ and $a \in \mathcal{A}$ and running all algorithms on the same set of potential rewards, whilst recording the regret incurred. We compare the performance of the LTS and OBS algorithms to that of UCB-Tuned, MOSS, UCB-V, KL-UCB and KL-UCB+ in each of the four simulated environments. The UCB-Tuned and MOSS algorithms are implemented exactly as described by Auer et al. (2002) and Audibert and Bubeck (2010) respectively.[3] The UCB-V algorithm is implemented as described by Audibert et al. (2007), with exploration function and tuning constants set to the 'natural values' suggested.[4] The KL-UCB and KL-UCB+ algorithms are implemented as described by Garivier and Cappé (2011), with constant $c = 0$, as used in their numerical experiments.

The results of the simulations are summarised in Figures 1–4. The left hand plots show cumulative regret averaged over the trials. The right hand plots show boxplots indicating the distribution of final cumulative regret over trials. We consider cumulative regret averaged over trials since this provides an estimate for the expected cumulative regret, $\mathbb{E}(R_T)$, where the expectation is taken with respect to the regressor sequence and the reward and action sequences under the proposed algorithm, a much more meaningful measure than the cumulative regret incurred over any one trial. We plot the average cumulative regret on a logarithmic timescale, so that one can get an indication as to whether an algorithm has a optimal rate of regret.

We first note that, in the cases considered, the MOSS and UCB-V algorithms perform relatively poorly, despite proven regret guarantees. The left hand plots in Figures 1 and 2 indicate that the KL-UCB+ algorithm has the best performance (in terms of expected regret) for the 'low variance' problem environments, whereas Figures 3 and 4 indicate that the UCB-Tuned algorithm has the best performance in the 'high variance' problem environments. Both the OBS and LTS algorithms display highly competitive performance in all cases considered, with the OBS algorithm consistently outperforming the LTS algorithm, as predicted in Section 1.2. It is also indicated that increasing the number of actions from 2 to 4 widens this performance gap between OBS and LTS. There

3. We implement the MOSS algorithm with the time horizon known. We note that the algorithm can be run without knowledge of the horizon using the 'doubling trick' (Cesa-Bianchi and Lugosi, 2006), whereby the horizon used in the algorithm is originally set to 2 and then doubled whenever $t$ exceeds the assumed horizon. In preliminary numerical experiments, the version using knowledge of the time horizon slightly outperformed (with respect to averaged final cumulative regret) the 'doubling trick' version in of all problem environments tested, so we choose to use the former in comparisons.
4. For the UCB-V algorithm, we use exploration function $\mathcal{E}_t = \log t$ and constant $c = 1/6$, in the notation of Audibert et al. (2007). In preliminary numerical experiments, this version outperformed the version used in the numerical experiments section of Audibert and Bubeck (2009) (with $c = 1$ instead) in all four problem environments tested, and so is used for comparisons.

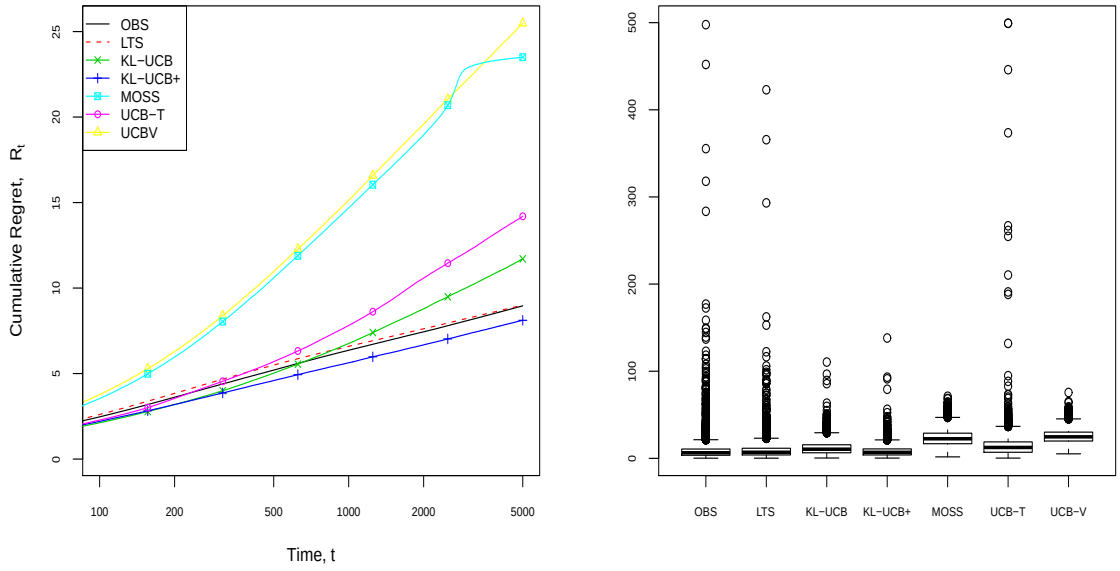


Figure 1: Performance of various algorithms in Bernoulli bandit simulation with parameters $(p_1, p_2) = (0.8, 0.9)$. Left: Cumulative regret averaged over trials. Right: Distribution of cumulative regret at time $t = 5000$. Results based on 8000 independent trials.

is no method tested that outperforms OBS in all four problems and the OBS algorithm displays performance that is never far from the leading algorithm.

The boxplots on the right hand side of the Figures 1–4 indicate that LTS, OBS, UBC-tuned and (to a lesser extent) KL-UCB+ are all 'risky' algorithms, when compared to the others. If one was risk-averse, then the KL-UCB, MOSS and UCBV algorithms are suitable options.[5] It is also worth noting that the regret distribution associated with the OBS algorithm seems to have a fatter upper tail than the LTS algorithm but the LTS algorithm has more variance near the median (which is higher than the OBS median in the four cases considered). A theoretical analysis on the concentration of regret for the OBS and LTS algorithms is desirable so that this can be investigated further, although we leave this to future work.

Finally, in Figure 5, we present plots of the reward ratio (2) through time, for the first 100 trials of the first experimental condition, in order to demonstrate actual results proved in the theoretical part of the paper. The 'almost sure' nature of the convergence of this quantity is observed, in that on some runs there is a period to begin with in which the ratio 'sticks' before asymptoting towards 1, whereas most runs converge quickly towards the asymptote. An identical phenomenon is observed in the other experimental conditions.

5. Note that Audibert and Bubeck (2009) give theoretical results on the concentration of the regret incurred by the UCB-V algorithm, as well as on its expectation.

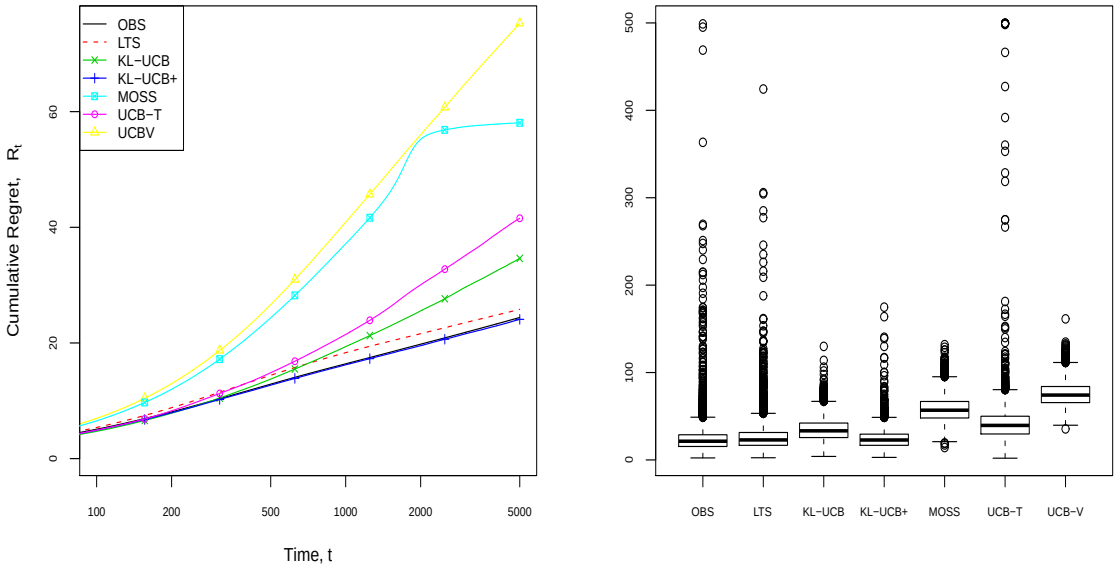


Figure 2: Performance in Bernoulli bandit simulation with parameters $(0.8, 0.8, 0.8, 0.9)$. Note that the curves for the OBS algorithm and the KL-UCB+ algorithm are virtually coincident.

### 4.2 Linear Regression

In this case, we study a form of the problem in which the expected reward for each action is a linear function of an observable scalar regressor and the reward noise terms are normally distributed. The learning task becomes that of estimating both the intercept and slope coefficients for each of the actions, so that the action yielding the highest expected reward given the regressor can be identified. The exploration-exploitation dilemma is inherent due to uncertainty in regression coefficient estimates caused by the reward noise.

#### 4.2.1 PROBLEM CONSIDERED

In this case, we let the prior information, $I_0$, consist of the following:

- The number of actions, $A = 4$.
- $(\forall t \in \mathcal{T})\{x_t \sim U(-0.5, 0.5)\}$.
- $(\forall a \in \mathcal{A})(\forall t \in \mathcal{T})\{f_a(x_t) = \beta_{1,a} + \beta_{2,a} x_t\}$ for $\beta_{1,a}, \beta_{2,a} \in \mathbb{R}$ unknown.
- $(\forall a \in \mathcal{A})(\forall t \in \mathcal{T})\{z_{t,a} \sim \mathrm{N}(0, \sigma_a^2)\}$ for $\sigma_a \in \mathbb{R}$ unknown.
- $(\forall a \in \mathcal{A})\{$The (improper) prior distributions for $\beta_{1,a}$ and $\beta_{2,a}$ are flat over $\mathbb{R}\}$.
- $(\forall a \in \mathcal{A})\{$The (improper) prior distribution of $\sigma_a^2$ is flat over $\mathbb{R}^+\}$.

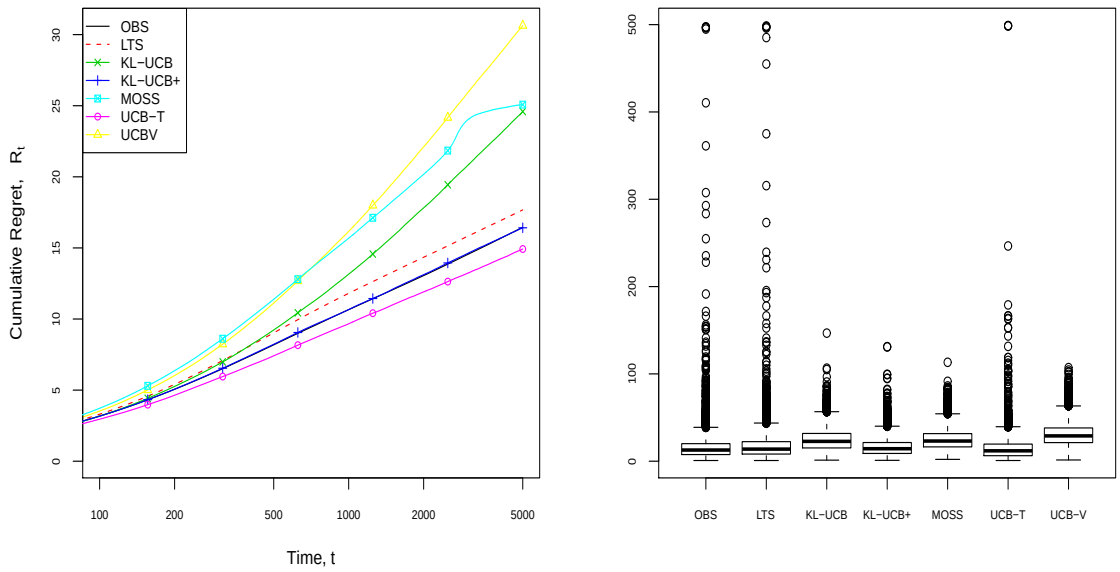


Figure 3: Performance in Bernoulli bandit simulation with parameters $(0.45, 0.55)$. Note that the curves for the OBS algorithm and the KL-UCB+ algorithm are virtually coincident.

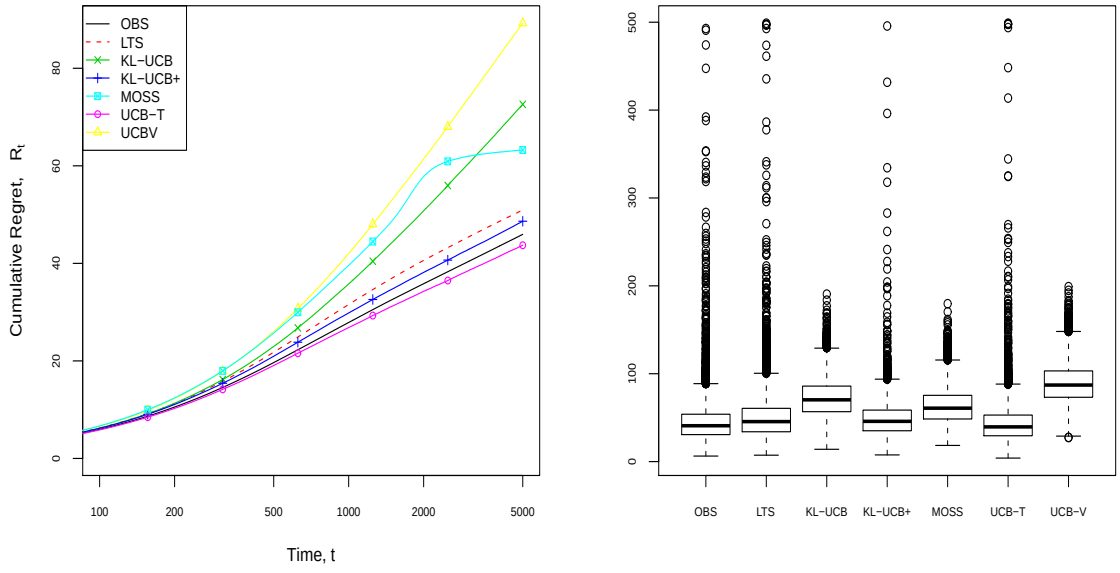


Figure 4: Performance in Bernoulli bandit simulation with parameters $(0.45, 0.45, 0.45, 0.55)$.

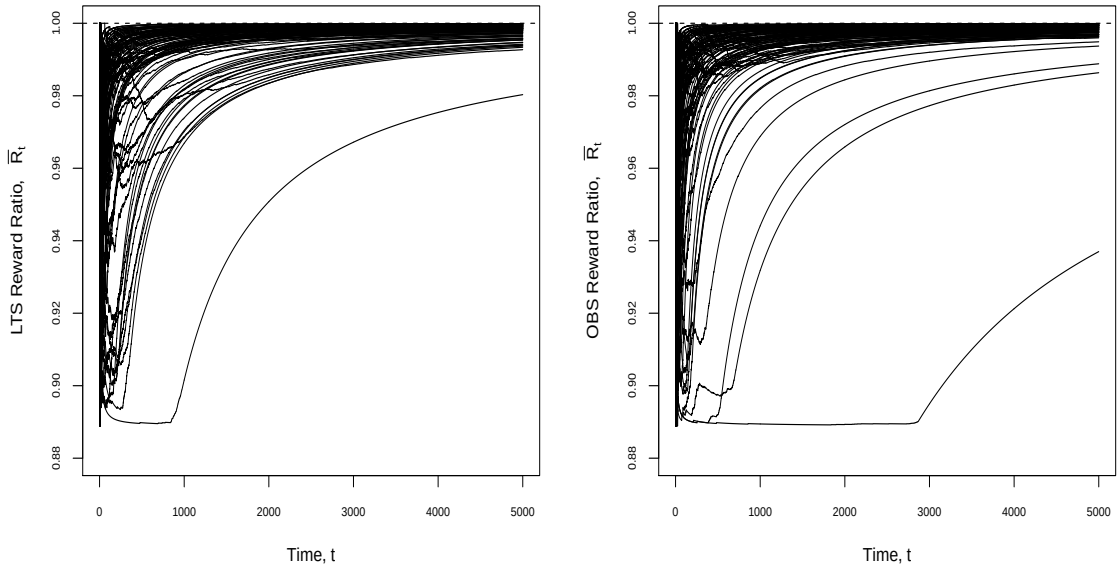


Figure 5: Convergence of the ratio (2) in the first 100 Bernoulli bandit simulations with parameters $(p_1, p_2) = (0.8, 0.9)$.

### 4.2.2 LTS AND OBS IMPLEMENTATION

Denote estimators at time $t$ of the parameters $\mathbf{b}_a$ and $\sigma_a$ for $a = 1, \ldots, A$ as $\hat{\mathbf{b}}_{t,a}$ and $\hat{\sigma}_{t,a}$ respectively, where $\mathbf{b}_a = (\beta_{1,a}, \beta_{2,a})^T$. For all $a \in \mathcal{A}$, denote $\mathcal{T}_{t,a} = \{\tau \in \{1, \ldots, t-1\} : a_\tau = a\}$ and the $n_{t,a}$-vectors of regressors and rewards observed at time steps in $\mathcal{T}_{t,a}$ as $\mathbf{x}_{t,a}$ and $\mathbf{r}_{t,a}$ respectively. Denote the $n_{t,a} \times 2$ matrix formed by the concatenation of $\mathbf{1}_{n_{t,a}}$ and $\mathbf{x}_{t,a}$ as $\mathbf{X}_{t,a}$, where $\mathbf{1}_{n_{t,a}}$ is the $n_{t,a}$-vector with every component equal to 1. Let $\hat{\mathbf{b}}_{t,a}$ be given by the least squares equation

$$\hat{\mathbf{b}}_{t,a} := (\mathbf{X}_{t,a}^T \mathbf{X}_{t,a})^{-1} \mathbf{X}_{t,a}^T \mathbf{r}_{t,a}.$$

Let us also denote $\mathbf{x}_t = (1, x_t)^T$. Posterior expectations (using flat priors, as indicated by $I_0$) can be evaluated easily, so we define exploitative value as

$$\hat{f}_{t,a}(x_t) := \mathbf{x}_t^T \hat{\mathbf{b}}_{t,a}.$$

Let $\hat{\sigma}_{t,a}$ be given by

$$\hat{\sigma}_{t,a} := \sqrt{\frac{1}{n_{t,a} - 2} (\mathbf{r}_{t,a} - \mathbf{X}_{t,a} \hat{\mathbf{b}}_{t,a})^T (\mathbf{r}_{t,a} - \mathbf{X}_{t,a} \hat{\mathbf{b}}_{t,a})}$$

and let $U_{t,a} \sim t_{n_{t,a}-2}$. We define the LTS exploratory value as

$$\tilde{f}_{t,a}^{\mathrm{Th}}(x_t) := \left[ \hat{\sigma}_{t,a} \sqrt{\mathbf{x}_t^T (\mathbf{X}_{t,a}^T \mathbf{X}_{t,a})^{-1} \mathbf{x}_t} \right] U_{t,a}. \tag{32}$$

The LTS total value is given by

$$Q^{\mathrm{Th}}_{t,a}(x_t) = \hat{f}_{t,a}(x_t) + \tilde{f}^{\mathrm{Th}}_{t,a}(x_t).$$

The OBS total value is given by

$$Q_{t,a}(x_t) = \max(Q^{\mathrm{Th}}_{t,a}, \hat{f}_{t,a}(x_t)).$$

Note that if $n_{t,a} \in \{0,1,2\}$ then the posterior distribution of $f_a(x_t)$ is improper. In these situations, we sample values from $N(0,10^3)$ to obtain $Q^{\mathrm{Th}}_{t,a}$.

#### 4.2.3 CONVERGENCE

In this section, we check explicitly that Assumptions 1–5 are satisfied in this linear regression setting, therefore proving that the LTS and OBS algorithms generate policies satisfying convergence criterion (2).

**Lemma 9** *The LTS total value estimate, $Q^{Th}_{t,a}$, satisfies Assumption 1, for all $a \in \mathcal{A}$.*

**Proof** Let $a \in \mathcal{A}$, $T > 0$, $I_T$ and $M \in \mathbb{R}$ be arbitrary. For any $t > T$ and $I_t = I_T \cup \{r_T, \ldots, r_{t-1}, a_T, \ldots, a_{t-1}\}$ with $a_s \neq a$ for $s \in \{T, \ldots, t-1\}$, the posterior distribution of $\mathbf{b}_a$ and $\sigma^2_a$ given $I_t$ will be the same as that given $I_T$ (since no further information about $f_a$ is contained in $I_t$). In particular for each regressor $x$, $\hat{f}_{t,a}(x) = \hat{f}_{T,a}(x)$, and $\tilde{f}^{\mathrm{Th}}_{t,a}(x)$ has the same distribution given $I_t$ as it did given $I_T$. Define

$$\varepsilon := \frac{1}{2} \min_{x \in [-0.5, 0.5]} \min\left\{ \mathbb{P}(\tilde{f}^{\mathrm{Th}}_{T,a}(x) < M - \hat{f}_{T,a}(x) \,|\, I_T), \mathbb{P}(\tilde{f}^{\mathrm{Th}}_{T,a}(x) < M - \hat{f}_{T,a}(x) \,|\, I_T) \right\}.$$

Since $Q^{\mathrm{Th}}_{t,a} = \hat{f}_{t,a}(x_t) + \tilde{f}^{\mathrm{Th}}_{t,a}(x_t)$, we then have that

$$\mathbb{P}(Q^{\mathrm{Th}}_{t,a} > M | I_t) > \varepsilon$$

and

$$\mathbb{P}(Q^{\mathrm{Th}}_{t,a} < M | I_t) > \varepsilon.$$

■

Lemma 10 (taken from Eicker, 1963) is used to prove the consistency of the least squares estimators of the regression coefficients.

**Lemma 10** *The least squares estimators $\hat{\mathbf{b}}_{t,a}$, $t = 2, 3, \ldots$ converge in probability to $\mathbf{b}_a$ as $n_{t,a} \to \infty$ if and only if $\lambda_{min}(\mathbf{X}^T_{t,a}\mathbf{X}_{t,a}) \to \infty$ as $n_{t,a} \to \infty$, where $\lambda_{min}(\mathbf{X}^T_{t,a}\mathbf{X}_{t,a})$ is the smallest eigenvalue of $\mathbf{X}^T_{t,a}\mathbf{X}_{t,a}$.*

**Lemma 11** *The exploitative value estimate $\hat{f}_{t,a}(x_t) \xrightarrow{\mathbb{P}} f_a(x_t)$ as $n_{t,a} \to \infty$.*

**Proof** Let $\tilde{x}_{i,a}$ denote the value of the regressor presented on the timestep where action $a$ was picked for the $i$th time.

$$\mathbf{X}^T_{t,a}\mathbf{X}_{t,a} = \begin{pmatrix} n_{t,a} & \sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a} \\ \sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a} & \sum_{i=1}^{n_{t,a}} \tilde{x}^2_{i,a} \end{pmatrix}.$$

The smallest eigenvalue is given by

$$\lambda_{\min} = \frac{n_{t,a}}{2}\left[\frac{\sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a}^2}{n_{t,a}} + 1 - \sqrt{\left(\frac{\sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a}^2}{n_{t,a}} + 1\right)^2 - 4\left(\frac{\sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a}^2}{n_{t,a}} - \frac{\left(\sum_{i=1}^{n_{t,a}} \tilde{x}_{i,a}\right)^2}{n_{t,a}^2}\right)}\right].$$

Therefore, since $\mathrm{Var}X > 0$, we have that

$$\lim_{n_{t,a}\to\infty} \lambda_{\min} = \lim_{n_{t,a}\to\infty} \frac{n_{t,a}}{2}\left[\mathbb{E}X^2 + 1 - \sqrt{\left(\mathbb{E}X^2+1\right)^2 - 4\mathrm{Var}X}\right] = \infty.$$

Using Lemma 10, we then have

$$\left[\hat{\mathbf{b}}_{t,a} - \mathbf{b}_a\right] \xrightarrow{\mathbb{P}} \mathbf{0} \text{ as } n_{t,a} \to \infty.$$

Multiplying on the left by $\mathbf{x}^T$ then gives

$$\left[\hat{f}_{t,a}(x_t) - f_a(x_t)\right] \xrightarrow{\mathbb{P}} 0 \text{ as } n_{t,a} \to \infty.$$

■

**Lemma 12** *The LTS total value estimate, $Q_{t,a}^{Th}$, satisfies Assumptions 2–4, for all $a \in \mathcal{A}$.*

**Proof** To prove this lemma, we need to show that $\tilde{f}_{t,a}(x_t) \xrightarrow{\mathbb{P}} 0$ as $n_{t,a} \to \infty$ for all actions $a \in \mathcal{A}$. In order to do this, we consider each component in the product that forms $\tilde{f}_{t,a}(x_t)$ (see (32)). Firstly, we consider $U_{t,a}$. It is a well known (as is described in Zwillinger, 2000) that

$$U_{t,a} \xrightarrow{\mathbb{D}} N(0,1) \text{ as } n_{t,a} \to \infty. \tag{33}$$

Next, we consider $\hat{\sigma}_{t,a}$. Using the facts that $\hat{\mathbf{b}}_{t,a} \xrightarrow{\mathbb{P}} \mathbf{b}_a$ as $n_{t,a} \to \infty$, $z_{t,a} = r_t - f_a(x_t)$ and $\mathbb{E}[z_{t,a}] = 0$ we have that

$$\begin{aligned}
\hat{\sigma}_{t,a} &:= \sqrt{\frac{1}{n_{t,a}-2}(\mathbf{r}_{t,a} - \mathbf{X}_{t,a}\hat{\mathbf{b}}_{t,a})^T(\mathbf{r}_{t,a} - \mathbf{X}_{t,a}\hat{\mathbf{b}}_{t,a})} \\
&\xrightarrow{\mathbb{P}} \sqrt{\frac{1}{n_{t,a}-2}(\mathbf{r}_{t,a} - \mathbf{X}_{t,a}\mathbf{b}_a)^T(\mathbf{r}_{t,a} - \mathbf{X}_{t,a}\mathbf{b}_a)} \text{ as } n_{t,a} \to \infty \\
&\xrightarrow{\text{a.s.}} \sqrt{\mathbb{E}[z_{t,a}^2]} \text{ as } n_{t,a} \to \infty \\
&= \sqrt{\mathrm{Var}[z_{t,a}] + [\mathbb{E}[z_{t,a}]]^2} \\
&= \sqrt{\mathrm{Var}[z_{t,a}]} = \sigma_a.
\end{aligned} \tag{34}$$

Finally, let us consider $\mathbf{x}_t^T(\mathbf{X}_{t,a}^T\mathbf{X}_{t,a})^{-1}\mathbf{x}_t$. We start by looking at the determinant of $X_{t,a}^T X_{t,a}$. We have that

$$\begin{aligned}\det X_{t,a}^T X_{t,a} &= n_{t,a}^2\left(\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}^2 - \Big(\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}\Big)^2\right)\\ &\overset{\text{a.s.}}{\to} n_{t,a}^2\Big(\mathbb{E}[x_t^2]-\mathbb{E}[x_t]^2\Big) \text{ as } n_{t,a}\to\infty\\ &= n_{t,a}^2\text{Var}[x_t]. \end{aligned} \tag{35}$$

Using the standard formula for inverting a $2\times 2$ matrix, we get

$$\begin{aligned}\mathbf{x}_t^T(\mathbf{X}_{t,a}^T\mathbf{X}_{t,a})^{-1}\mathbf{x}_t =& \frac{1}{\det X_{t,a}^T X_{t,a}}\left(\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}^2 - 2x_t\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a} + x_t^2 n_{t,a}\right)\\ =& \frac{1}{\det X_{t,a}^T X_{t,a}}\left(n_{t,a}\Big[\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}^2 - \Big(\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}\Big)^2\Big]\right.\\ &\left.+\frac{1}{n_{t,a}}\Big(\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}\Big)^2 - 2x_t\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a} + x_t^2 n_{t,a}\right)\\ =& \frac{1}{n_{t,a}} + \frac{1}{\det X_{t,a}^T X_{t,a}}\left(\frac{1}{n_{t,a}}\Big(\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}\Big)^2 - 2x_t\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a} + x_t^2 n_{t,a}\right)\\ =& \frac{1}{n_{t,a}} + \frac{1}{\det X_{t,a}^T X_{t,a}}\left(n_{t,a}\Big[\Big(\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a}\Big)^2 - 2\frac{1}{n_{t,a}}x_t\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a} + x_t^2\Big]\right)\\ =& \frac{1}{n_{t,a}} + \frac{1}{\det X_{t,a}^T X_{t,a}}\left(n_{t,a}\Big[\frac{1}{n_{t,a}}\sum_{i=1}^{n_{t,a}}\tilde{x}_{i,a} - x_t\Big]^2\right). \end{aligned} \tag{36}$$

Using (35), (36) and the facts that $\text{Var}[x_t]>0$ and both $x_t$ and $\mathbb{E}[x_t]$ are bounded, we have that

$$\mathbf{x}_t^T(\mathbf{X}_{t,a}^T\mathbf{X}_{t,a})^{-1}\mathbf{x}_t \overset{\text{a.s.}}{\to} \frac{1}{n_{t,a}} + \frac{\Big[\mathbb{E}[x_t]-x_t\Big]^2}{n_{t,a}\text{Var}[x_t]} \overset{\text{a.s.}}{\to} 0 \text{ as } n_{t,a}\to\infty. \tag{37}$$

Equations (33), (34) and (37) imply that $\tilde{f}_{t,a}^{\text{Th}}(x_t)\overset{\mathbb{P}}{\to}0$ as $n_{t,a}\to\infty$. Therefore, since $Q_{t,a}^{\text{Th}} = \hat{f}_{t,a}(x_t) + \tilde{f}_{t,a}^{\text{Th}}(x_t)$, Lemma 11 gives us that

$$Q_{t,a}^{\text{Th}} - f_a(x_t)\to 0 \text{ as } n_{t,a}\to\infty,$$

satisfying Assumptions 2 and 4. This same holds for $\hat{f}_{t,a}(x_t)$, hence Assumption 3 is satisfied too. ■

**Lemma 13** *The exploitative value estimate, $\hat{f}_{t,a}(x_t)$, satisfies Assumption 5, for all $a\in\mathcal{A}$.*

**Proof** Let $a\in\mathcal{A}$, and $T>0$ be arbitrary. For any $t>T$ and $I_t = I_T\cup\{r_T,\ldots,r_{t-1}, a_T,\ldots,a_{t-1}\}$ with $a_s\neq a$ for $s\in\{T,\ldots,t-1\}$, the regression coefficients $\hat{\mathbf{b}}_{t,a}$ are equal to $\hat{\mathbf{b}}_{T,a}$. Hence

$$\max_{x\in[-0.5,0.5]}\hat{f}_{t,a}(x) = \max_{x\in[-0.5,0.5]}\hat{f}_{T,a}(x).$$

The Assumption then follows by noting that $\mathcal{S} = \mathbb{R}$. ■

**Proposition 2** *Within the described linear regression setting convergence criterion (2) is satisfied when the LTS or the OBS algorithm is used.*

**Proof** Assumptions 1–5 hold, so the proof follows directly from Theorems 1 and 2. ■

#### 4.2.4 EXPERIMENTAL RESULTS

The process is run for 10000 independent trials. A time window of $\mathcal{T} = \{1, \ldots, 5000\}$ is considered on each trial. The regression coefficients for the actions are set to $(\beta_{1,a}, \beta_{2,a}) = (0,1), (0,-1), (-0.1,0), (0.1,0)$ for $a = 1,2,3,4$ respectively. The resulting expected reward functions are plotted in Figure 6. For each trial:

- $\forall t \in \mathcal{T}$ sample $x_t \sim \mathrm{U}(-0.5, 0.5)$
- $\forall a \in \mathcal{A}$ and $\forall t \in \mathcal{T}$ sample $z_{t,a} \sim \mathrm{N}(0, \sigma_a^2)$ with $\sigma_a = 0.5$
- $\forall a \in \mathcal{A}$ and $\forall t \in \mathcal{T}$ evaluate potential reward $r_{t,a} = \beta_{1,a} + \beta_{2,a} x_t + z_{t,a}$
- record the regret incurred using various action selection methods.

We compare the performance of LTS and OBS to an interval estimation method (or LinUCB, in the terminology of Li et al., 2010) similar to that described in Pavlidis et al. (2008). However we use the posterior distribution of the mean to evaluate the upper confidence bound rather than using the predictive distribution. Specifically, the action selection rule used is given by

$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left[ \hat{f}_{t,a}(x_t) + \hat{\sigma}_{t,a} \sqrt{\mathbf{x}_t^T (\mathbf{X}_{t,a}^T \mathbf{X}_{t,a})^{-1} \mathbf{x}_t} \right] t_{1-\frac{\lambda}{100}, n_{t,a}-2}$$

where $t_{\gamma,n}$ denotes the quantile function of Student's $t$ distribution with $n$ degrees of freedom evaluated at $\gamma$. This ensures that the value estimates are consistent, that is, the value estimates converge to the true expected reward as associated actions are selected infinitely often. We implement the IE method with parameter values $\lambda = 0.01$, $\lambda = 5$ and $\lambda = 25$.

The results of the simulation can be seen in Figures 7 and 8. Figure 7 (left) shows cumulative regret averaged over the trials. The OBS algorithm displays the best performance (with respect to cumulative regret averaged over trials) in the problem considered, and this performance is significantly better than that of the LTS algorithm. It is also clear that the IE method performance is highly sensitive to parameter choice. The best parameter choice in this case is $\lambda = 5$, however, it is not clear how this parameter should be chosen based on the prior information provided. In general, if $\lambda$ is 'too high', then too much emphasis is put on short term performance and if $\lambda$ is 'too low' then too much emphasis is put on long term performance. This is indicated by the curves for the $\lambda = 25$ and $\lambda = 0.01$ methods respectively. Figure 7 (right) shows boxplots indicating the distribution of final cumulative regret over trials. It is indicated that the IE methods become riskier as the significance parameter used is increased and that the significance parameter provides a way of trading off median efficiency and risk. The only method to compete with OBS on cumulative regret averaged

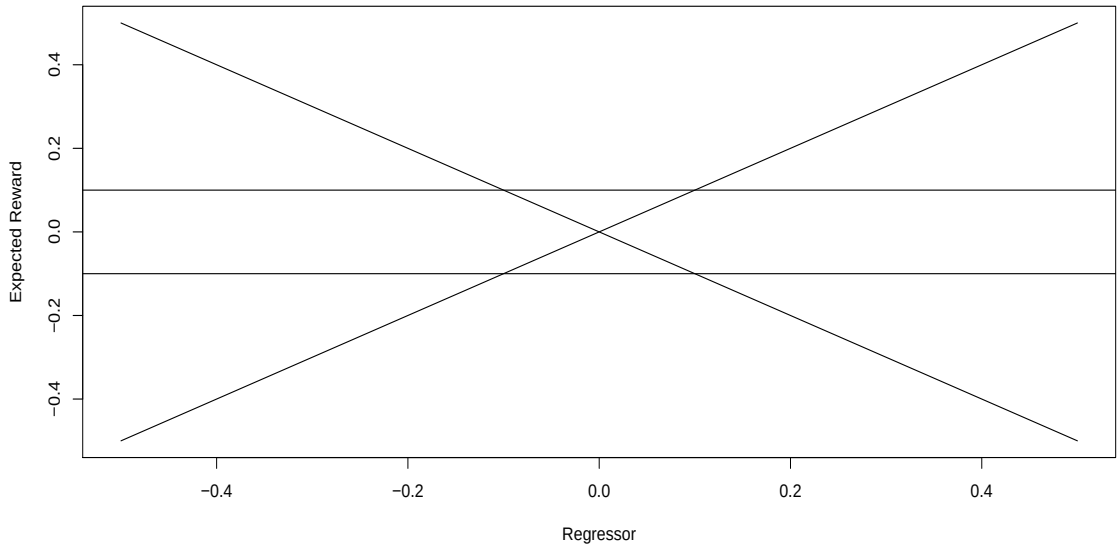


Figure 6: The expected reward functions for the 4 actions in linear regression simulation.

over trials is the $\lambda = 5$ IE method, however the OBS final regret distribution is more concentrated than the $\lambda = 5$ IE method. In Figure 8, we present plots of the reward ratio (2) through time, for the first 100 experiments, in order to demonstrate actual results proved in the theoretical part of the paper. Although convergence of the ratio has not occurred after the 5000 iterations, it is clear that the ratio is improving over time.

### 4.3 Web-Based Personalised News Article Recommendation

We now consider the problem of selecting news articles to recommend to internet users based on information about the users. In our framework, the recommendation choice corresponds to an action selection and the user information corresponds to a regressor. The objective is to recommend an article that has the highest probability of being clicked.

We test the performance of the LTS and OBS algorithms on a real-world data set, the Yahoo! Front Page Today Module User Click Log Data Set (Yahoo! Academic Relations, 2011). A similar study is performed by Chapelle and Li (2011). However we consider multiple trials over a short time horizon, as opposed to Chapelle and Li's single trial over the full data set, to investigate the short term performance of the algorithms, and in particular to address the claim made in Section 1.2 regarding a potential short term benefit of using OBS over using LTS. It is necessary to average results over multiple trials given the randomised nature of the OBS and LTS algorithms. We also test the LinUCB algorithm of Li et al. (2010) with various parameter settings to provide a benchmark for comparison.

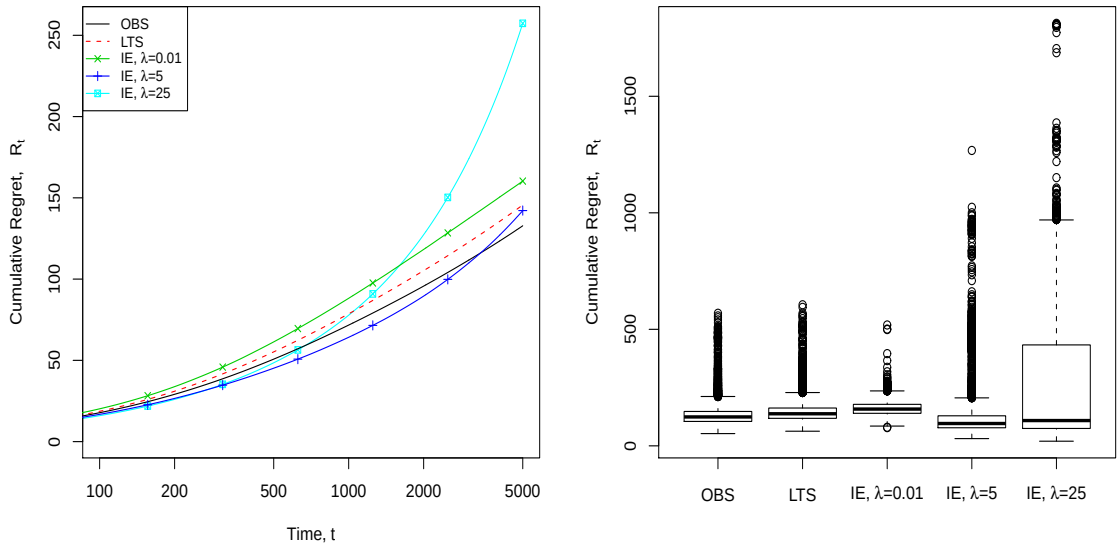


Figure 7: Performance of various algorithms in linear regression simulations. Left: Cumulative regret averaged over trials. Right: Distribution of cumulative regret at $t = 5000$. Results based on 10000 independent trials.

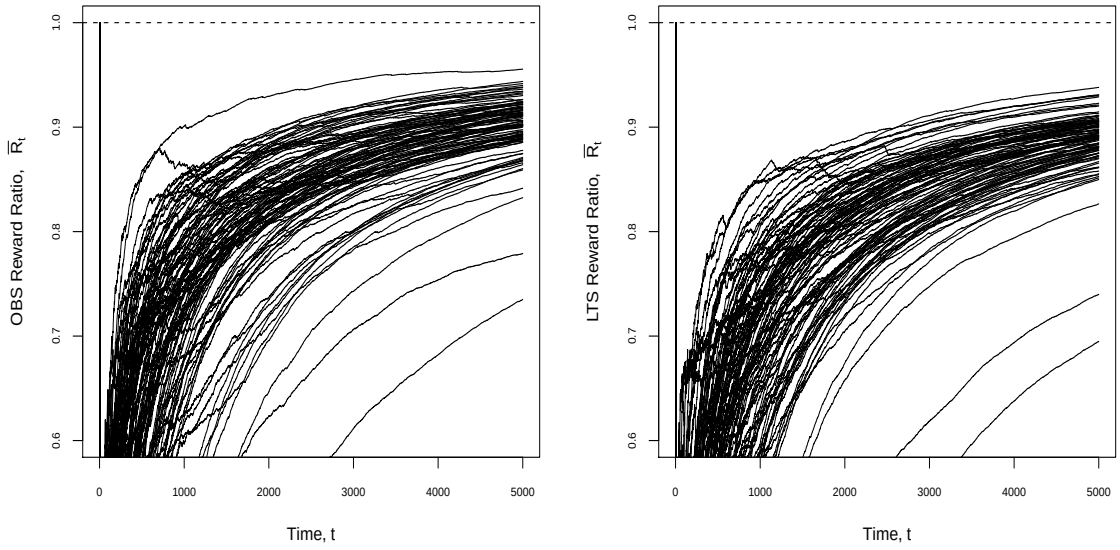


Figure 8: Convergence of the ratio (2) in the first 100 linear regression simulations.

#### 4.3.1 USE OF DATA SET

The data set describes approximately 36M instances of news articles being recommended to internet users on the Yahoo! Front Page Today Module at random times in May 2009. The form and collection of the data set are both described in detail by Li et al. (2010). For each recommendation, the data contains information concerning which article was recommended, whether the recommendation was clicked and a feature vector describing the user. The recommended articles are chosen uniformly at random from a dynamic pool of about 20 choices, with articles being added and removed at various points of the process. The user features, $\mathbf{x}_t$, are given as vectors of length 6 with one component fixed to 1, and are constructed as described by Li et al. (2010). The reward is defined to be 1 if the recommendation is clicked and 0 otherwise.

The use of past data presents a problem in evaluating a decision-making algorithm. Specifically, within the data a random article is recommended on each instance, which might well be different to the article that the decision-making algorithm selects during testing. This problem can be avoided by implementing the unbiased offline evaluator procedure of Li et al. (2011). Under this procedure, if the action selected by the algorithm does not match the action selected in the data point, the current data point, and subsequent data points, are ignored until a data point which matches user data and action selection occurs. The observed reward from this data point is then awarded to the algorithm, and the user data from the next recommendation instance in the data is used in the next evaluation step.

#### 4.3.2 ALGORITHM IMPLEMENTATION

The LTS and OBS algorithms are implemented using the logistic regression model of Chapelle and Li (2011). It is assumed that there is an unknown weight vector, $\mathbf{w}_a$, for each article $a \in \mathcal{A}$ such that

$$\mathbb{P}(r_t = 1 | a_t = a, \mathbf{x}_t = \mathbf{x}) = (1 + \exp(-\mathbf{w}_a^T \mathbf{x}))^{-1}.$$

Approximate posterior distributions for each $\mathbf{w}_a$ are estimated to be Gaussian with mean and variance updates as described in Algorithm 3 of Chapelle and Li (2011). For our numerical experiment, we set the unspecified regularisation parameter of Chapelle and Li (2011) to 100. The LTS algorithm can easily be implemented by sampling weight vectors from the posteriors and selecting the article with the weight vector forming the highest scalar product with the current user feature vector. The OBS algorithm can easily be implemented by also considering posterior means of these scalar products. We also test the LinUCB algorithm, as implemented by Chapelle and Li (2011), with parameter $\alpha$ set to each of 0.5, 1 and 2.

#### 4.3.3 NUMERICAL EXPERIMENTS

As previously mentioned, our focus is short term performance averaged over numerous trials. We focus on the case of only 4 articles, and therefore remove all instances outwith these 4 articles from the data set. On each of $2,500$ trials, we run each of the 5 algorithms until $5,000$ interactions are accepted using data from the start of the supplied data set (Yahoo! Academic Relations, 2011); we use only data from the start of the data set to avoid confounding the algorithm evaluations with the non-stationarity of the data.

The concept of regret is difficult to use as a performance measure in this setting, since there is no true model given for comparison. We instead consider the percentage of past timesteps resulting in clicks, otherwise known as the click-through rate (CTR), and percentage benefit of OBS over LTS

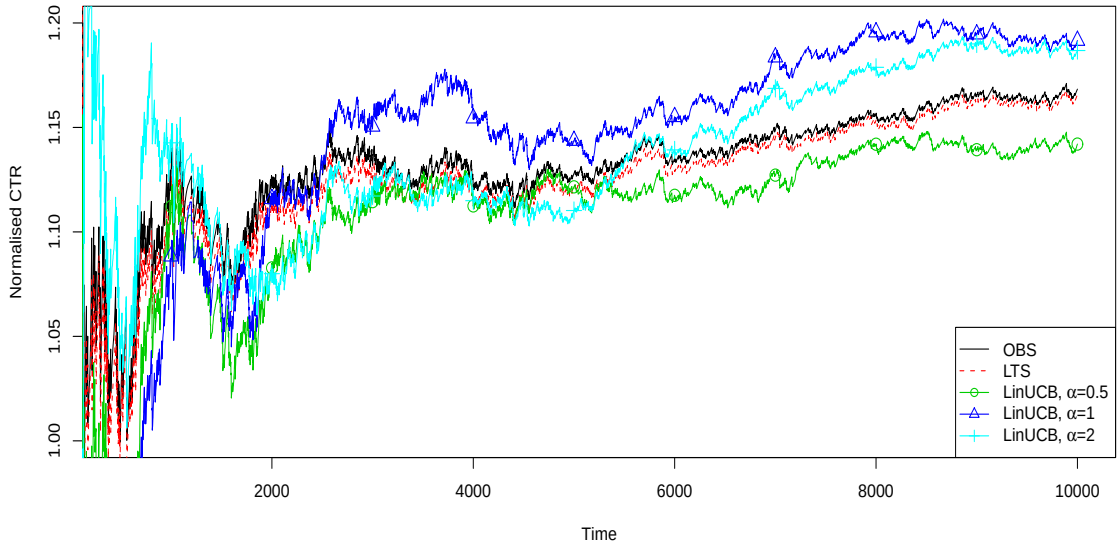


Figure 9: Normalised Click Through Rate through time for various algorithms. Results averaged over 2,500 independent trials.

with respect to CTR. Again, to avoid issues of non-stationarity, we normalise all CTRs by dividing by the CTR achieved (on these four articles) in the original data set.

The results of the experiment can be found in Figures 9 and 10. Figure 9 shows the normalised CTR for all 5 algorithms, averaged over all 2500 runs. It is clear that the performance of the LinUCB algorithm is sensitive to parameter choice; the version with parameter set to 1 performs much better than the version set to 0.5, and it is not clear in advance of implementing the algorithm which parameter will be optimal. As a caveat on these results, it is worth noting that the portion of the data set used for each trial is the same, and also that the LinUCB algorithms are deterministic given past information (except in the case of a tie in action values), so it is hard to extrapolate general results relating to the performance of LinUCB algorithms. Furthermore Chapelle and Li (2011) explain that the performance of the LinUCB algorithm degrades significantly with increasing feedback delay, while the LTS and OBS algorithms are more robust to the delay, so the strong performance of the highest-performing LinUCB algorithm in this experiment should not be taken as conclusive evidence of high real-world performance. Unfortunately it is not possible to produce plots comparable to Figures 5 and 8 in this case since the true optimal actions are not known. Figure 10 shows the difference in performance of OBS and LTS, expressed as a percentage of LTS performance, averaged over all 2500 runs. It is clear that the OBS algorithm outperforms the LTS algorithm across the time period considered, validating the intuition in Section 1.2. The short term improvement is small, but in many web-based application, a small difference in performance can be significant (Graepel et al., 2010).

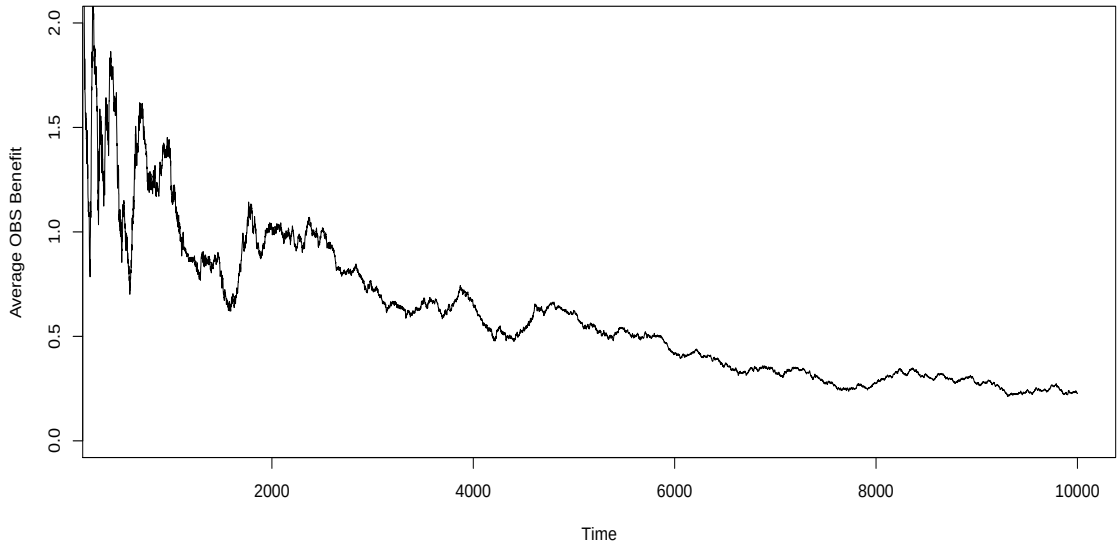


Figure 10: OBS CTR as a percentage improvement of LTS CTR through time. Results averaged over $2,500$ independent trials.

## 5. Discussion

The assumptions made for the theoretical results in Section 3 are mild in the sense that one would expect them to hold if the true posterior distributions and expectations are used. It is worth noting that convergence criterion (2) is satisfied even when approximations to the posterior distributions and expectations for the $f_a(x_t)$ are used with the LTS and OBS algorithms, so long as the relevant assumptions are satisfied. Hence, convergence is guaranteed for a large class of algorithms.

We have seen that both the LTS and the OBS algorithms are easy to implement in the cases considered. They are also computationally cheap and robust to the use of posterior approximations, when compared to belief-lookahead methods, such as Gittins indices. The simulation results for the OBS algorithm are very encouraging. In every case, the OBS algorithm outperformed the LTS algorithm and performed well compared to recent benchmarks.

## Acknowledgments

We thank the reviewers and editor for their comments, which have contributed significantly to the quality of this article. This research was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Networks) project and is jointly funded by a BAE Systems and EPRSC (Engineering and Physical Sciences Research Council) strategic partnership (EP/C548051/1). Many of the ideas in this paper grew out of extensive discussions with Nicos Pavlidis and Niall Adams.

## References

R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. arXiv:1111.1797v1, 2011.

J. Asmuth, L. Li, M. L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proc. 25th Conf. on Uncertainty in Artificial Intelligence*, pages 19–26, Corvallis, Oregon, 2009. AUAI Press.

J.-Y. Audibert and S. Bubeck. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 150–165. Springer, 2007.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

M. Brezzi and T.L. Lai. Incomplete learning from endogenous data in dynamic allocation. *Econometrica*, 68:1511–1516, 2000.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *25th Annu. Conf. on Neural Information Processing Systems*, NIPS 2011, Granada, Spain, 2011.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *14th Internat. Conf. on Artificial Intelligence and Statistics*, AISTATS, Fort Lauderdale, Florida, USA, 2011.

F. Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, 34:447–456, 1963.

S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *24th Annu. Conf. on Neural Information Processing Systems*, pages 586–594. Curran Associates, Inc., 2010.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *24th Annu. Conf. on Learning Theory*, COLT 2011, Budapest, Hungary, 2011.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, B*, 41:148–177, 1979.

T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proc. of the 27th Internat. Conf. on Machine Learning*, pages 13–20, Haifa, Israel, 2010. Omnipress.

O.-C. Granmo. A Bayesian learning automaton for solving two-armed bernoulli bandit problems. In *7th Internat. Conf. on Machine Learning and Applications*, San Diego, California, USA, 2008. IEEE Computer Society.

L. P. Kaelbling. Associative reinforcement learning: Functions in k-DNF. *Machine Learning*, 15: 279–298, 1994.

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

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc. 19th Internat. Conf. on World Wide Web*, WWW 2010, pages 661–670, Raleigh, North Carolina, USA, 2010. ACM.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proc. 4th ACM Internat. Conf. on Web Search and Data Mining*, WSDM '11, pages 297–306, 2011.

M. L. Littman. A generalized reinforcement-learning model: Convergence and applications. In *Proc. 13th Internat. Conf. on Machine Learning*, pages 310–318. Morgan Kaufmann, 1996.

N. Meuleau and P. Bourgine. Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35:117–154, 1999.

N. G. Pavlidis, D. K. Tasoulis, and D. J. Hand. Simulation studies of multi-armed bandits with covariates. In *Proc. 10th Internat. Conf. on Computer Modelling*, Cambridge, UK, 2008.

S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 39:287–308, 2000.

A. Slivkins. Contextual bandits with similarity information. arXiv:0907.3986v3, 2011.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

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

T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *Proc. 22nd Internat. Conf. on Machine Learning*, pages 956–963, New York, NY, USA, 2005. ACM.

Yahoo! Academic Relations. Yahoo! front page today module user click log dataset, version 1.0, 2011. URL `http://webscope.sandbox.yahoo.com`.

Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, 30:100–121, 2002.

D. Zwillinger. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press, 2000.