Apple Tasting Revisited: Bayesian Approaches to Partially Monitored Online Binary Classification

James A. Grant*[1] and David S. Leslie†[1]

[1]Department of Mathematics and Statistics, Lancaster University, UK

September 29, 2021

Abstract

We consider a variant of online binary classification where a learner sequentially assigns labels (0 or 1) to items with unknown true class. If, but only if, the learner chooses label 1 they immediately observe the true label of the item. The learner faces a trade-off between short-term classification accuracy and long-term information gain. This problem has previously been studied under the name of the 'apple tasting' problem. We revisit this problem as a partial monitoring problem with side information, and focus on the case where item features are linked to true classes via a logistic regression model. Our principal contribution is a study of the performance of Thompson Sampling (TS) for this problem. Using recently developed information-theoretic tools, we show that TS achieves a Bayesian regret bound of an improved order to previous approaches. Further, we experimentally verify that efficient approximations to TS and Information Directed Sampling via Pólya-Gamma augmentation have superior empirical performance to existing methods.

Keywords: Partial Monitoring, Binary Classification, Thompson Sampling, Information-Directed Sampling, Pólya-Gamma Augmentation

1 Introduction

Binary classification is a fundamental problem in statistics, machine learning, and many applications. Its online version, wherein a learner iteratively guesses the classes of items and has their true classes revealed has also been well-studied (e.g. Angluin, 1988; Littlestone, 1990; Cesa-Bianchi et al., 1996; Ying and Zhou, 2006). In this paper, we study a variant of the stochastic online binary classification where the true class is *only* revealed for items guessed to belong to a particular fixed class, irrespective of whether that guess is correct.

Such a problem has previously been studied under the name of the 'Apple Tasting Problem', inspired by a toy problem of learning to visually identify bad or rotten apples in a packaging plant (Helmbold et al., 1992, 2000). In this example, a learner considers apples one by one and makes a choice whether or not to taste the apple based on its appearance - apples are either good or

*j.grant@lancaster.ac.uk; corresponding author
†d.leslie@lancaster.ac.uk

bad and their quality can, to some extent, be determined by their appearance. The learner wishes to minimise the number of good apples tasted (as a tasted apple cannot be sold), and maximise the number of bad apples tasted (as this prevents bad produce going to market - the learner is willing to sacrifice their own taste buds for profit). The difficulty of the problem enters through the aspects that the learner does not know the function relating appearance to quality and only receives information on the quality of the apple if they taste it.

Apples aside, conceptually similar problems may arise in numerous settings. For instance, in quality control settings products arrive to a decision-maker sequentially. Products may be satisfactory or faulty, and the decision-maker can determine this by investigating the product, subject to a further cost. It may be prohibitively costly to investigate every product but it is also important to identify and remove faulty products, to avoid greater costs further down the line. To balance their costs the decision-maker must make sufficiently many investigations to learn the characteristics of faulty products.

As another example, companies monitoring credit card fraud sequentially observe transactions, a small proportion of which are fraud. There is effectively no cost to allowing a non-fraudulent transaction to proceed, but costs are associated with investigating a transaction (whether fraudulent or not) and with allowing a fraudulent transaction to proceed. The natural aim is to allow genuine transactions to continue unimpeded, and investigate the fraudulent transactions, but immediate information on the class of a transaction can only be gathered through investigation.

More broadly, problems of this flavour can arise in a range of online monitoring and supervised learning settings: such as filtering spam emails (Sculley, 2007), classifying traffic on networks, and label prediction in many contexts. An understanding of optimal algorithms for these sequential classification tasks is therefore valuable to decision-makers in a wide range of applications. The next section introduces a mathematical framework for the apple tasting problem.

1.1 Logistic Contextual Apple Tasting

We consider an online binary classification task taking place over T rounds. In each round $t \in [T] \equiv \{1, \dots, T\}$, an item arrives with associated feature vector $x_t \in \mathcal{X} \subset \mathbb{R}^d$. This item has a latent class $C_t \in \{0, 1\}$, which is modelled as having a stochastic dependence on the feature vector x_t, and an unknown parameter vector $\theta^* \in \Theta \subset \mathbb{R}^d$. Note that θ^* does not vary with t, and that Θ is known. Throughout this paper, the distribution of the class in round t is given as

$$C_t \mid x_t \sim Bern(\sigma(x_t^\top \theta^*)),$$

where σ denotes the logistic function[1], such that $\sigma(z) = (1 + e^{-z})^{-1}$, $z \in \mathbb{R}$.

Also in round t, an agent, who we call 'the learner', guesses the class of the item with feature vector x_t. The guessed class is represented by $A_t \in \{0, 1\}$. If $A_t = 1$, (a loss indicative of) the true class C_t is revealed. If $A_t = 0$, no feedback is received. It is known that if the true class is $C_t = 0$, choosing $A_t = 0$ incurs zero loss, and $A_t = 1$ incurs a loss of one. If $C_t = 1$, choosing $A_t = 1$ incurs loss $l_{11} \geq 0$, which represents the cost of intervention, and $A_t = 0$, incurs cost $l_{01} \geq l_{11}$.

The expected loss of an action $a \in \{0, 1\}$ in round t with respect to parameter θ is then given

[1]The nature of our work is such that analogous theoretical results should be readily available for other suitably smooth transforms, e.g. the probit function.

by the function $\mu_t : \{0,1\} \times \Theta \to \mathbb{R}$ defined as,

$$\mu_t(a,\theta) := \begin{cases} l_{01}\sigma\left(x_t^\top \theta\right), & a = 0, \\ 1 + (l_{11} - 1)\sigma\left(x_t^\top \theta\right), & a = 1. \end{cases} \tag{1}$$

We define the optimal action in round t w.r.t. a parameter $\theta \in \Theta$ as

$$\alpha_t(\theta) \in \operatorname*{argmin}_{a \in \{0,1\}} \mu_t(a,\theta),$$

and let $A_t^* = \alpha_t(\theta^*)$, i.e. A_t^* is the action optimal with respect to the true parameter θ^* in round t.

We are interested in the expected performance of the learner, given they use a particular rule (or *policy*) to assign labels. Formally, a policy φ is a mapping from a history[2],

$$\mathcal{H}_{t-1} = \varsigma(A_1, x_1, C_1 A_1, \ldots, A_{t-1}, x_{t-1}, C_{t-1} A_{t-1}, x_t),$$

to an action $a_t \in \{0,1\}$ for any $t \in \mathbb{N}$. The learner's performance is captured by the *Bayesian regret* of the policy φ in T rounds,

$$BR(T,\varphi) = \mathbb{E}_0 \left(\sum_{t=1}^{T} \mu_t\left(A_t, \theta^*\right) - \mu_t\left(A_t^*, \theta^*\right) \right),$$

where this expectation is taken with respect to a prior, π_0, on θ^* supported on Θ.

The Bayesian regret measures the difference between the expected loss of an oracle decision maker who knows the status distribution of each item and acts optimally with respect to this information, and the expected loss of the learner making decisions according to the policy φ. We will be interested in the scaling of the Bayesian regret with respect to T and to the dimension d of the context vectors.

We will principally be interested in the performance of the Thompson Sampling (TS) policy. In round t, TS plays the action $\alpha_t(\theta_t)$ where θ_t is drawn from the posterior distribution $\pi_{t-1} = \pi_0(\cdot \mid \mathcal{H}_{t-1})$.

The idea behind TS - of choosing actions according to the current posterior belief - dates back as far as Thompson (1933). However it is in the last decade that it has become popular as a general approach for sequential decision-making problems, and has been shown to achieve optimal or near-optimal scaling of regret across many settings including multi-armed bandits (Agrawal and Goyal, 2013a), contextual bandits (May et al., 2012; Agrawal and Goyal, 2013b), structured bandits (Russo and Van Roy, 2016; Grant and Leslie, 2020), and a linear variant of finite partial monitoring (Tsuchiya et al., 2020).

1.2 Partial Monitoring

To fully characterise our logistic apple tasting problem and its best achievable regret, it is useful to cast the problem as part of the broader partial monitoring (PM) framework. The apple tasting problem is one of the simplest examples of a PM problem. Specifically, our proposed model is an example of PM with side information, as considered in Bartók and Szepesvári (2012).

[2] We use ς here to denote the sigma-algebra to avoid overloading notation. Notice, in particular that $\mathcal{H}_{t-1}$ includes the round t feature vector x_t.

It can also be shown that subject to a rescaling of the loss matrix (Antos et al., 2013; Lienert, 2013), our framework coincides with a logistic contextual bandit (Filippi et al., 2010) with only two actions in each round (one always being a zero vector). The present problem however is a very specific instance of this much broader framework, and a bespoke treatment of apple tasting is therefore useful as a complement to existing theory for the much more general setting (Dong et al., 2019; Faury et al., 2020).

A general PM problem with side information is formalised as a tuple $\mathbf{G} = (\mathbf{L}, \mathbf{\Phi}, \mathcal{F})$, where $\mathbf{L} \in \mathbb{R}^{N \times M}$ is a loss matrix, $\mathbf{\Phi} \in \Sigma^{N \times M}$ is a feedback matrix[3], and $\mathcal{F}$ is a set of possible mappings from contexts to outcome distributions. In our case,

$$\mathbf{L} = \begin{pmatrix} 0 & l_{01} \\ 1 & l_{11} \end{pmatrix} \text{ and } \mathbf{\Phi} = \begin{pmatrix} 0 & 0 \\ 1 & l_{11} \end{pmatrix},$$

and $\mathcal{F}$ is the set of logistic functions parameterised by $\theta \in \Theta$.

In each of a series of rounds $t \in [T]$, a context $x_t \in \mathcal{X}$ is drawn (not necessarily at random). The learner observes the context and then chooses an action $n_t \in [N]$. An outcome $m_t \in [M]$ is drawn from the distribution $f(x_t)$ and the learner receives loss $\mathbf{L}_{n_t, m_t}$ and observes feedback $\mathbf{\Phi}_{n_t, m_t}$. In our problem, n_t corresponds to the choice of label, m_t is the true class, and the losses and feedbacks are the corresponding entries of $\mathbf{L}$ and $\mathbf{\Phi}$.

The best achievable regret in PM problems is well understood in a minimax sense. In particular, for both stochastic (Bartók et al., 2011, 2014) and adversarial (Bartók et al., 2010; Antos et al., 2013) finite PM (where M and N are finite), it has been shown that all games can be classified as either 'trivial', 'easy', 'hard', or 'hopeless' based on $(\mathbf{L}, \mathbf{\Phi})$, and have associated minimax regret of order $\Theta(1)$, $\Theta(\sqrt{T})$, $\Theta(T^{2/3})$, or $\Theta(T)$ respectively.

In both settings, the apple tasting problem is shown to belong to the class of 'easy' problems, with $\Theta(\sqrt{T})$ minimax regret[4]. It is worth noting that the term *easy* refers to the learnability of the problem, but does not imply that the design of an optimal algorithm is trivial.

For a general family of easy problems with side information (including apple tasting) Bartók and Szepesvári (2012) prove that an upper-confidence-bound-based algorithm, CBP-SIDE, realises this $\Theta(\sqrt{T})$ regret. Furthermore, Lienert (2013) shows empirically that both CBP-SIDE and the LinUCB algorithm (Li et al., 2010; Chu et al., 2011) for contextual bandits, are effective for a linear (as opposed to logistic) variant of our problem.

However, in practice, upper-confidence-bound based approaches can be overly conservative, and only behave competitively when T is very large. In this paper we show that an improved empirical performance can be achieved by randomised, Bayesian algorithms such as TS, without sacrificing theoretical guarantees.

1.3 Related Literature

As mentioned previously, our principal focus in this paper is the performance of TS - Russo et al. (2018) gives a detailed summary of developments around TS across many problems. Our analysis of the Bayesian regret is based on a recent line of information-theoretic analysis (Russo and Van Roy, 2016; Dong and Van Roy, 2018; Dong et al., 2019; Lattimore and Szepesvári, 2019) which has been

[3] Here Σ is some known, finite alphabet.

[4] Note that these bounds are in the frequentist setting, but such bounds automatically imply equivalent results in the Bayesian setting (the converse is not true). See e.g. Section 34.7 of Lattimore and Szepesvári (2020)

shown to be useful for problems with large context or action sets. In utilising these ideas in the apple tasting setting, we extend a number of existing results for finite parameter sets to compact $\Theta \subset \mathbb{R}^d$. An alternative strategy for bounding the Bayesian regret, not considered here, has been to exploit frequentist confidence sets to construct high probability guarantees on which actions are selected (Russo and Van Roy, 2014b; Grant et al., 2019; Grant and Leslie, 2020).

A related algorithm, inspired by the link between information gain and the necessary exploration in sequential decision-making problems is Information-Directed Sampling (IDS), introduced in Russo and Van Roy (2014a, 2018). Like TS, IDS also selects at random based on the posterior belief, but constructs the distribution from which this sample is drawn based on a trade-off of expected regret, and expected information gain given the feedback on the upcoming action. IDS (and frequentist approximations thereof) has been applied to certain bandit, partial monitoring, and reinforcement learning problems (Liu et al., 2018; Kirschner and Krause, 2018; Kirschner et al., 2020a,b; Arumugam and Van Roy, 2020) and shown strong empirical and theoretical results, comprable to those for TS. We discuss the use of IDS for apple tasting in Section 4, and evaluate it empirically alongside TS in Section 5.

This paper is the first work we aware of that specifically applies TS to apple tasting, but previous work has considered its use for logistic bandits. For logistic contextual bandits, the implementation of exact TS (i.e. the policy that draws its sample from the exact posterior) is infeasible due to the intractability of the posterior distribution. It is therefore necessary to sample from an approximation of the posterior to implement a TS-like approach. Dumitrascu et al. (2018) recently proposed an approximation based on Polya-Gamma augmentation (Polson et al., 2013; Windle et al., 2014) which has improved convergence properties over Laplace approximation originally used by Chapelle and Li (2011). We investigate such a Polya-Gamma augmentation-based approximation in the context of apple tasting in Section 3. The effect of approximation of the posterior on the performance of TS is an area of increasing interest, as Phan et al. (2019) have recently proved that a small constant approximation error can induce linear regret in the application of TS to certain simple multi-armed bandit problems. Appropriately designed approximate algorithms can be successful however, as shown theoretically (Mazumdar et al., 2020) for particular Langevin approximation algorithms, and empirically in a range of settings (e.g. Urteaga and Wiggins, 2018).

The apple tasting problem is not the only variant of online classification where labels are not revealed in every round. In selective classification (or classification with a reject (or abstention) option) (e.g. Chow, 1957; Sayedi et al., 2010; Wiener and El-Yaniv, 2011) the learner may decline to label items, thus mitigating the risk of labelling when they have high uncertainty. Conversely, in classification with selective sampling (Cesa-Bianchi et al., 2009; Orabona and Cesa-Bianchi, 2011; Cavallanti et al., 2011; Dekel et al., 2012; Agarwal, 2013), the learner must label all items, but observing labels is costly, and the learner has the option to decline to observe the label if it is deemed to have insufficient informational value. The same problem has also been studied under the name 'online active binary classification' (Monteleoni and Kaariainen, 2007; Liu et al., 2015). Both of these variants differ from apple tasting in that they have a more complex action set.

Gentile and Orabona (2014) consider a multi-class label prediction problem where the learner chooses a subset of possible labels, in each round, and only observes true labels if they are part of their subset. While this also lies at the intersection of classification and partial monitoring, when the number of classes is reduced to two, i.e. in the binary classification setting, this problem reduces to the usual full-feedback online problem. Apple tasting is therefore more challenging in the 2-class setting due to the information imbalance between the actions.

1.4 Motivations and Contributions

Our motivations for a renewed treatment of apple tasting are threefold. Firstly, despite the existence of alternative theoretically justified approaches, new developments in the theoretical understanding of TS allow us to derive guarantees for the empirically superior TS policy. Second, the apple tasting setting strikes a sufficient balance between simplicity and complexity to allow an uncluttered study of the effect of zero-information actions in PM, and of posterior approximation to the performance of TS. Finally, as outlined in the first section, the range of applications of the apple tasting problem are broad, and our empirical investigation of methods such as TS that have been popularised with the last decade is therefore useful and pertinent.

The principal contributions of our work are the following:

- We provide an information-theoretic analysis of the Bayesian regret of TS for logistic contextual apple tasting (LCAT). This gives rise to an $\tilde{O}(\sqrt{dT})$ bound on regret which is optimal with respect to horizon T (up to logarithmic factors), and sharper with respect to the feature dimension d than the best frequentist bounds for UCB-type algorithms. Notably, the bound is also of an improved order with respect to d than the $\tilde{O}(d\sqrt{T})$ bounds achievable in the more general contextual bandit setting. Our analysis extends theoretical techniques previously only used for finite parameter spaces to the more readily modelled setting of compact but infinite parameter spaces.

- We adapt the Pólya-Gamma TS scheme of Dumitrascu et al. (2018) to give approximate TS and information directed sampling (IDS) schemes appropriate to LCAT. In the TS setting, we show that this scheme is an asymptotically consistent approximation to exact TS.

- We identify a potential issue in the application of IDS to contextual problems, or those with non-stationary expected information gains. Namely, that it may fail to take account of the magnitude of the information gain and make counter-intuitive selections as a result. We explain how a tunable variant avoids this issue.

- We validate the efficacy of the Pólya-Gamma-based TS and tunable-IDS by showing their superior performance to competitor approaches on simulated data.

Our theoretical results on TS are given in Section 2. In Section 3 we discuss the Polya-Gamma augmentation scheme necessary for a practical implementation of TS, and in Section 4 we discuss the limitations of a similar implementation of IDS. Finally, we demonstrate the efficacy of TS numerically in Section 5.

2 Bayesian Regret of Thompson Sampling

Our first theoretical result is given in this section. It bounds the Bayesian regret of TS under any sequence of bounded feature vectors. To complete our formalisation of the setting in which theoretical guarantees can be established, we suppose w.l.o.g. that the parameter set Θ lies in the unit ball in $\mathbb{R}^d$, i.e. $\Theta \subset B_1^d$, and that there exists a bound $x_{\max} < \infty$ on the dimensions of feature vectors, i.e. $||x||_\infty \leq x_{\max}$, for all $x \in \mathcal{X}$.

Theorem 1 *For the contextual logistic apple tasting problem instantiated by $\theta^* \sim \pi_0$, the Bayesian regret of the Thompson Sampling policy, φ^{TS}, in T rounds satisfies,*

$$BR\left(T, \varphi^{TS}\right) = O\left(\sqrt{dT\log(T)}\right). \tag{2}$$

In the following section, 2.1, we introduce some information theoretic concepts which are required in the proof of the bound, which is given in Section 2.2.

The result in Theorem 1 may imply the stronger performance of TS than alternative approaches. The CBP-SIDE algorithm of Bartók and Szepesvári (2012) achieves a near-optimal frequentist regret, but still with an order greater than our Bayesian regret bound:

$$Reg\left(T, \varphi^{CBP-SIDE}\right) = O\left(d^2\log(T)\sqrt{T}\right).$$

In particular, its dependence on the dimension of the parameter vector is worse by a factor of $d^{3/2}$. Casting the problem as a contextual logistic bandit, as outlined in Antos et al. (2013); Lienert (2013) and directly utilising the existing results of Dong et al. (2019) for said more general setting gives an $\tilde{O}(d\sqrt{T})$ bound on Bayesian regret. Our difference of a factor of $\sqrt{d}$ in (2) is a result of our bespoke analysis for the 2-action apple tasting setting.

The other previously existing approaches for contextual apple tasting of Helmbold et al. (2000), which transform binary classification algorithms to apple tasting algorithms, guarantee a regret that is sublinear in T but that is linear in the size of the context set $|\mathcal{X}|$. Such bounds are therefore not useful in the present setting with infinitely many possible contexts.

2.1 Preliminaries

Before giving the proof of Theorem 1, we require some additional notation and concepts. Firstly, in-keeping with the notation for more general PM problems, we define the incurred loss and observed signal as $L_t = L(C_t, A_t)$ and $\Phi(A_t) = L_t\mathbb{I}\{A_t = 1\}$.

Our bound will rely on information-theoretic techniques, and as such a definition of the mutual information between probability distributions is necessary. For random variables X, and Y, following distributions P_X, and P_Y respectively, define the mutual information between X and Y as

$$I(X;Y) = D_{KL}(P_{X,Y}||P_X \otimes P_Y),$$

where D_{KL} denotes the Kullback-Leibler divergence, and $P_{X,Y}$ denotes the joint distribution of X and Y.

Related to this, a key quantity will be the expected information gained by the learner about the parameter θ^* in a single round. To define this, let $\mathbb{E}_t$ denote expectation with respect to the posterior density π_t induced by the history $\mathcal{H}_t$, and introduce I_t as a function giving the mutual information with respect to this posterior. For random variables X and Y adapted to π_{t-1} define

$$I_t(X;Y) = I\left(X;\ Y \mid \mathcal{H}_{t-1}\right).$$

The expected information gained about θ^* in a single round t by the learner using TS is then represented by $I_t(\theta^*; (\theta_t, \Phi_t(\alpha_t(\theta_t))))$.

This expected information gain plays a key role in the following quantity called the *information ratio*. For any Θ-valued random variables θ, θ' and $t \in [T]$ we define the information ratio as

$$\Gamma_t(\theta, \theta') = \frac{\left[\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\theta'), \theta^*\right) - \mu_t\left(\alpha_t(\theta), \theta^*\right)\right)\right]^2}{I_t\left(\theta; (\theta', \Phi_t(\alpha_t(\theta')))\right)}. \tag{3}$$

When $\theta = \theta^*$ this is the ratio of the square of the expected regret incurred in round t by a decision-maker acting as though θ' is the true parameter and the expected information gained about θ^* as a result of their decision. Thus a large information ratio corresponds to high regret and low expected information, whereas a small information ratio corresponds to low regret and high expected information.

The information ratio is a quantity, introduced in a more general setting by Russo and Van Roy (2016), which allows for a useful decomposition of the Bayesian regret of Thompson Sampling and related randomised approaches. When the information ratio can be uniformly bounded for all $t \in [T]$ several studies have successfully derived order-optimal bounds on the Bayesian regret in various settings (Russo and Van Roy, 2016, 2018; Dong et al., 2019; Lattimore and Szepesvári, 2019). In such settings, the bound is expressed terms of such a uniform bound on $\Gamma_t(\cdot, \cdot)$, and the entropy of the distribution on θ^*.

Here however, the realisation of such a bound would not be finite as the parameter space Θ is not discrete, and thus $H(\theta^*)$ lacks a finite bound. Fortunately, this does not mean that the problem of bounding the regret in our setting is hopeless. To achieve sublinear regret it is not necessary to learn the distribution over all of Θ. Following Dong and Van Roy (2018) we introduce a *rate distortion* of the parameter θ^* to learn a simpler ϵ-optimal action function rather than the exact optimal action function.

For $\theta, \theta' \in \Theta$, and a round $t \in [T]$ we define the *distortion rate* as

$$d_t(\theta, \theta') = \mu(\alpha_t(\theta), \theta') - \mu(\alpha_t(\theta'), \theta').$$

This measures the difference in the expected loss computed with respect to θ' of the optimal action given θ and θ'. It happens to coincide with the regret of TS when $\theta' = \theta^*$, the true parameter, and $\theta = \theta_t$, the sample drawn by TS in round t. Further, define $\{\Theta_k\}_{k=1}^K$ to be a partition of Θ in K parts such that, for a fixed and arbitrary $\epsilon > 0$

$$d_t(\theta, \theta') \leq \epsilon, \quad \forall \theta, \theta' \in \Theta_k, \ \forall t \in [T], \quad \text{for any } k \in [K].$$

Then define an associated indexing random variable ϕ_ϵ on $[K]$ such that

$$\phi_\epsilon = k \Leftrightarrow \theta^* \in \Theta_k. \tag{4}$$

This variable indicates which cell of the partition the true parameter θ^* lies in. The entropy of ϕ_ϵ, $H(\phi_\epsilon)$ is bounded by $\log(K)$. Thus if the structure of Θ permits a small K, $H(\phi_\epsilon)$ can be much smaller than $H(\theta^*)$. In what follows, we will derive a bound on the Bayesian regret of TS that depends on $H(\phi_\epsilon)$.

2.2 Proof of Theorem 1

The regret upper bound arises as a result of approximating the regret of TS using the posterior on θ^* with the regret of TS using a discrete distribution. The basic premise is to relate the regret to

that of an algorithm that is expected to incur at least $o(\epsilon)$ regret in each round due to discretisation, and then study the additional regret incurred in the simpler discrete setting. The bound of the desired order is then obtained by specifying ϵ as a decreasing function of T. We emphasise that ϵ and the discretised variant of TS are purely hypothetical constructs for the proof, and are not required for the actual implementation of TS - thus the decision-maker does not have to worry about identifying and tuning them.

The first step in our proof constructs a series of discrete random variables $\tilde{\theta}_t^*$, $t \in [T]$ which approximate θ^*. These random variables are functions of θ^* whose realisations lie in the same cell of the partition as θ^*. However, their dependence on θ^* can be expressed entirely via the variable ϕ_ϵ, such that $\tilde{\theta}_t^*$ is independent of θ^* conditioned on ϕ_ϵ. The following proposition asserts the existence of these random variables, and verifies their certain key properties. This result extends a similar result (Proposition 2 of Dong and Van Roy (2018)) to the setting where Θ is not a discrete set. Below, the random variable $\tilde{\theta}_t$, which has the same marginal distribution as $\tilde{\theta}_t^*$, can be thought of as a discretised analog of the Thompson sample θ_t.

Proposition 1 *For ϕ_ϵ as defined in (4), there exists a random variable $\tilde{\theta}_t^*$, supported on up to $2K$ points in Θ, in each round t satisfying the following properties:*

(i) Conditioned on ϕ_ϵ, $\tilde{\theta}_t^*$ is independent of θ^*,

(ii) $\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\theta_t), \theta^*\right) - \mu_t\left(\alpha_t(\theta^*), \theta^*\right)\right) \leq \epsilon + \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right)$, *a.s.,*

(iii) $I_{t-1}\left(\phi_\epsilon; (\tilde{\theta}_t, \Phi(\alpha_t(\tilde{\theta}_t)))\right) \leq I_{t-1}\left(\phi_\epsilon; (\theta_t, \Phi(\alpha_t(\theta_t)))\right)$, *a.s.,*

where in (ii) and (iii), $\tilde{\theta}_t$ is independent from and identically distributed to $\tilde{\theta}_t^*$.

Properties (ii) and (iii) say that the distribution of $\tilde{\theta}_t$ is such that the extra regret incurred by following TS using $\tilde{\theta}_t$ is no more than ϵ, and that the information gain about the compressed random variable ϕ_ϵ is not more than that gained using TS. The proof of Proposition 1 is given in Appendix A.

We use the properties of $\tilde{\theta}_t^*$, and Γ_t to decompose the expected regret in a single round. Essentially, the steps below bound per-round-regret by accepting a constant regret of ϵ, to move from considering regret of TS to the regret of the discretised TS. For the regret in round t, $\Delta_t = \mu\left(\alpha_t(\theta_t), \theta^*\right) - \mu\left(\alpha_t(\theta^*), \theta^*\right)$, we have,

$$\begin{aligned}
\mathbb{E}_{t-1}(\Delta_t) &= \mathbb{E}_{t-1}\left(\mu\left(\alpha_t(\theta_t), \theta^*\right) - \mu\left(\alpha_t(\theta^*), \theta^*\right)\right) \\
&\leq \epsilon + \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right) && \text{by (ii)} \\
&= \epsilon + \sqrt{\Gamma_t(\tilde{\theta}_t^*, \tilde{\theta}_t) I_t\left(\tilde{\theta}_t^*; \left(\tilde{\theta}_t, \Phi(\alpha_t(\tilde{\theta}_t))\right)\right)} && \text{by definition of } \Gamma_t \\
&\leq \epsilon + \sqrt{\Gamma_t(\tilde{\theta}_t^*, \tilde{\theta}_t) I_t\left(\phi_\epsilon; \left(\tilde{\theta}_t, \Phi(\alpha_t(\tilde{\theta}_t))\right)\right)} && \text{data processing inequality} \\
&\leq \epsilon + \sqrt{\Gamma_t(\tilde{\theta}_t^*, \tilde{\theta}_t) I_t\left(\phi_\epsilon; (\theta_t, \Phi(\alpha_t(\theta_t)))\right)} && \text{by (iii)} \qquad (5)
\end{aligned}$$

The result is that the regret in a single round has been decomposed in terms of the information ratio, and expected information gain. We proceed to bound this further, via a uniform bound on the information ratio, as given by the following proposition.

Proposition 2 *There exists a constant $\bar{\Gamma} > 0$ such that for every round $t \in \mathbb{N}$*

$$\Gamma_t(\tilde{\theta}_t^*, \tilde{\theta}_t) \leq \frac{\bar{\Gamma}}{\pi_{t-1}(\Theta_t)}, \tag{6}$$

where $\Theta_t := \{\theta \in \Theta : \alpha_t(\theta) = 1\}$ is the set of parameters such that the revealing label is chosen, and $\pi_{t-1}(\Theta_t) = \int_{\theta \in \Theta_t} \mathrm{d}\pi_{t-1}(\theta)$ is the posterior mass placed on this set.

Similar to Proposition 1, the proof of this Proposition extends ideas from Dong and Van Roy (2018) to non-finite Θ, and the partial monitoring setting. The techniques used in this extension could also be used to extend to non-finite parameter spaces in other problems. Principally, the proof consists of lower bounding the information gain via Pinsker's inequality, and relating the expected loss function to the sigmoid function to bound the information ratio. The full proof is provided in Appendix C.

It follows from Proposition 2 that the per-round regret can be bounded as follows,

$$\begin{aligned}
\mathbb{E}_{t-1}(\Delta_t) &\leq \epsilon + \sqrt{\frac{\bar{\Gamma}}{\pi_{t-1}(\Theta_t)} I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(\alpha(\theta_t)))\right)} && \text{by (5) and (6)} \\
&= \epsilon + \sqrt{\frac{\bar{\Gamma}}{\pi_{t-1}(\Theta_t)} \pi_{t-1}(\Theta_t) I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(1))\right)} \\
&= \epsilon + \sqrt{\bar{\Gamma} I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(1))\right)}, && (7)
\end{aligned}$$

where the first equality is true because if $A_t = 0$, there is no information gain. Now, aggregating the regret over T rounds, we have

$$\begin{aligned}
BR(T) &= \sum_{t=1}^{T} \mathbb{E}_{t-1}(\Delta_t) && \text{Tower Rule} \\
&\leq T\epsilon + \sum_{t=1}^{T} \sqrt{\bar{\Gamma} I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(1))\right)} && \text{by (7)} \\
&\leq T\epsilon + \sqrt{T\bar{\Gamma} \sum_{t=1}^{T} I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(1))\right)} && \text{by Cauchy-Schwarz} \\
&\leq T\epsilon + \sqrt{T\bar{\Gamma} H(\phi_\epsilon)}, && (8)
\end{aligned}$$

with the final inequality holding since the expected information gained about ϕ_ϵ is ultimately bounded by its entropy, i.e.

$$\sum_{t=1}^{T} I_t\left(\phi_\epsilon; (\theta_t, \Phi_t(1))\right) \leq I(\phi_\epsilon; \theta^*) \leq H(\phi_\epsilon).$$

The final step in the proof is to bound the $H(\phi_\epsilon)$ term in (8). We recall that $H(\phi_\epsilon) \leq \log(K)$, where K is the size of the partition of Θ. The following proposition bounds K, and has its proof in Appendix D.

Proposition 3 *There exists a partition* $\{\Theta_k\}_{k=1}^K$ *of* Θ*, satisfying*

$$d_t(\theta, \theta') \leq \epsilon, \ \theta, \theta' \in \Theta_k, \ \forall k \in [K], \ \forall t \in [T], \tag{9}$$

for any $\epsilon \in (0, \sigma(\delta) - 0.5)$*, such that*

$$K \leq \left(\frac{3 l_{max} x_{max}}{\epsilon} \right)^d. \tag{10}$$

The proof is completed by combining (8) and (10) and choosing $\epsilon = O\left(T^{-1/2}\right)$. □

3 Pólya-Gamma Thompson Sampling

The logistic classification model is such that the posterior on θ^*, π_t, for any $t \geq 1$ is intractable. This renders the implementation of TS via samples from the exact posterior in each round infeasible, and necessitates the use of samples from an approximate posterior.

In the related setting of the logistic contextual bandit, Dumitrascu et al. (2018) introduce an approximate variant of TS which uses Pólya-Gamma (PG) augmentation to admit efficient sampling. We adapt this to give an approximate TS policy for LCAT, PG-TS, in Algorithm 1. It utilises a Gibbs sampler for the unknown parameters, possible due to a parameter augmentation approach, and highly efficient rejection sampler for the augmenting parameters due to Polson et al. (2013); Windle et al. (2014). A full description of the Gibbs sampler is provided in Appendix E.

In our algorithms we let the function $\texttt{GIBBS}(\mathbf{b}, \mathbf{B}, M, \mathcal{D}, \theta)$ denote the use of said Gibbs sampler initialised at parameter θ to draw M samples from the approximation of the posterior implied by a prior, $MVN_\Theta(\mathbf{b}, \mathbf{B})$ which is a Gaussian with mean $\mathbf{b}$ and covariance $\mathbf{B}$, restricted to Θ, and observed data $\mathcal{D}$.

Inputs: Prior mean vector $\mathbf{b}$, Prior covariance matrix $\mathbf{B}$, Number of Gibbs iterations M.

Initialise: $\mathcal{D} = \emptyset$, and $\theta_0^{(M)} \sim MVN_\Theta(\mathbf{b}, \mathbf{B})$.
for $t = 1, 2, \ldots$ **do**
  $\{\theta_t^{(1)}, \ldots, \theta_t^{(M)}\} \leftarrow \texttt{GIBBS}(\mathbf{b}, \mathbf{B}, M, \mathcal{D}, \theta_{t-1}^{(M)})$
  Receive context $x_t \in \mathbb{R}^d$
  Select action $A_t = \alpha_t(\theta_t^{(M)})$
  if $A_t = 1$ **do**
    Observe $\Phi_t(1) \in \{l_{01}, l_{11}\}$
    Augment $\mathcal{D} \leftarrow \mathcal{D} \cup \{x_t, \Phi_t(1)\}$
  end if
end for

Algorithm 1: PG-TS

3.1 On the Impact of Approximate Inference

The regret results of the Section 2 are based on exact sampling from the posterior. The PG-TS algorithm necessarily samples from an approximation of the posterior, to maintain a reasonable

computational overhead. Recent work of Phan et al. (2019) has identified conditions under which sampling from an approximate posterior can lead to linear regret in multi-armed bandit problems. On the other hand, May et al. (2012) have shown that sublinear regret in contextual bandits can be achieved without drawing samples from an *exact* posterior - in fact that it suffices to sample from a distribution that converges around the true parameters in the limit.

In this section we address the possible concern around use of an approximate sampler, by demonstrating that PG-TS does not meet the sufficient conditions identified by Phan et al. (2019), and further adapt the results of May et al. (2012) to LCAT to show that PG-TS obtains an asymptotically sublinear regret.

Phan et al. (2019) characterise approximate TS policies in terms of their α-divergence[5], defined for a pair of distributions P, Q with densities $p(x)$, $q(x)$, and a coefficient $\alpha \in \mathbb{R}$ as,

$$D_\alpha(P, Q) = \frac{1 - \int p(x)^\alpha (1 - q(x))^{1-\alpha} \mathrm{d}x}{\alpha(1-\alpha)}. \tag{11}$$

The α-divergence generalises a number of divergences including $KL(Q, P)$ (when $\alpha \to 0$) and $KL(P, Q)$ (when $\alpha \to 1$), and can be related to the Total Variation (TV) distance via Pinsker's inequality.

At a high level, the contribution of Phan et al. (2019) is to show that there exist approximate distributions Q_t which satisfy $D_\alpha(\Pi_t, Q_t) < \epsilon$ for a true posteriors Π_t and constant $\epsilon > 0$, but sampling from Q_t at every time step results in linear regret. In effect this shows that $D_\alpha(\Pi_t, Q_t) < \epsilon$ is not a sufficient condition for approximate TS according to Q_t to inherit sublinear regret guarantees of exact TS according to Π_t.

In our setting then, let θ_t be the sample used in round t under exact TS, and $\theta_t^{(M)}$ be the sample used in round t under PG-TS. Let π_t, and $\pi_t^{(M)}$ be the densities associated with these random variables, assuming the same truncated multivariate Gaussian prior in both cases. Our first theoretical result of this section, Theorem 2 below, will demonstrate that in the limit (with respect to t) the α-divergence between π_t and $\pi_t^{(M)}$ goes to 0. Therefore the PG-TS scheme is not unreliable in the sense identified by Phan et al. (2019).

Our second result, Theorem 3 below, uses the results of May et al. (2012) to show that the expected regret of PG-TS is indeed asymptotically sublinear in T. Theorem 1 of May et al. (2012) establishes sufficient conditions for asymptotic consistency of a randomised contextual bandit algorithm. Specifically, they consider the contextual bandit problem where context x_t in a closed $\mathcal{X}$ is observed, and action a_t in a finite set $\mathcal{A}$ is selected, inducing reward observation $r_t = f_{a_t}(x_t) + z_{t,a_t}$, where $f_a : \mathcal{X} \to \mathbb{R}$ are unknown continuous functions, and $z_{t,a}$ are zero-mean random variables. May et al. (2012) establish conditions under which the sequence of chosen actions $\{a_1, a_2, \dots\}$ satisfies the following convergence criterion of Yang and Zhu (2002),

$$\frac{\sum_{t=1}^T f_{a_t}(x_t)}{\sum_{t=1}^T f^*(x_t)} \to 1 \text{ a.s., as } t \to \infty, \tag{12}$$

where $f^*(x) = \text{argmax}_{a \in \mathcal{A}} f_a(x)$ is the optimal expected reward.

Both theorems make use of additional assumptions on the parameter space and context distribution, which require the following additional notation. For $\theta \in \mathbb{R}^d$ define the sets $\mathcal{X}_1(\theta) = \{x \in$

[5]Here we use the script α to avoid confusion with α_t, the optimal action selection function.

$\mathcal{X} : \alpha(x, \theta) = 1\}$, and $\mathcal{X}_0(\theta) = \mathcal{X} \setminus \mathcal{X}_1(\theta)$, of context vectors such that actions 1 and 0, respectively, are optimal.

The following assumptions on the properties of $\mathcal{X}_1$ and $\mathcal{X}_0$ ensure that our approximate posterior distributions converge appropriately. Assumption 1 ensures repeated sampling, regardless of where the posterior mass initially gathers, and Assumption 2 ensures that some items of class 0 will be observed among the revealed true labels.

Assumption 1 *Contexts are drawn i.i.d. from distribution p_X on $\mathcal{X}$, and there exists $\delta > 0$ such that $p_X(\mathcal{X}_1(\theta))$, $p_X(\mathcal{X}_0(\theta)) > \delta$ for every $\theta \in \Theta$.*

Assumption 2 *For every $\theta \in \Theta$, there exists $x \in \mathcal{X}_1(\theta)$ such that $\sigma(x^\top \theta^*) < 1$.*

We are now ready to present our theoretical results relating to the performance of PG-TS. The proofs of the theorems are given in Appendices F and G respectively.

Theorem 2 *Under Assumptions 1 and 2, and for the densities π_t and $\pi_t^{(M)}$ as defined above, we have*

$$\lim_{T \to \infty} D_{KL}\left(\pi_T, \pi_T^{(M)}\right) = 0. \tag{13}$$

Theorem 3 *Under Assumptions 1, and 2, the sequence of actions $\{A_t^{(M)}\}_{t=1}^{\infty}$ selected by PG-TS satisfies the convergence criterion (12), and thus the regret of PG-TS is asymptotically sublinear.*

To go further than these results - i.e. to establish that the regret guarantees associated with exact TS carry to PG-TS - is likely to be much more complex. The most advanced results on the regret of approximate TS in simple multi-armed bandits, for instance, rely on complex Bayesian non-parametric theory, which, to the best of our knowledge, does not yet have a known analog applicable to contextual problems (Mazumdar et al., 2020).

4 Information Directed Sampling

The PG-augmentation scheme can also be used to devise an approximate Information Directed Sampling (IDS) scheme, based on the framework proposed by Russo and Van Roy (2018). IDS algorithms are randomised policies which construct an action sampling distribution, in each round t, based on a trade-off of regret and information gain. They have been shown to enjoy a similar, or improved, theoretical and empirical performance to TS as well as a potential for generalisation to a wider range of partial monitoring problems, since they do not restrict themselves to selecting actions which have a non-zero probability of being optimal. Approximate IDS schemes can also be realised via the same PG-augmentation scheme as used for PG-TS, as described in this section.

Two general methods to select the IDS action sampling distribution have been proposed. Both are designed to trade-off between achieving a low expected regret, and a high expected information gain. We shall explain these in a more general bandit-type loss-minimisation setting where $\mathcal{A}_t \subseteq \mathcal{A} \subset \mathbb{R}^d$ denotes a (potentially continuous) action set at time t, $l : \mathcal{A} \to \mathbb{R}$ denotes an expected loss function, and $\Delta_t : \mathcal{A} \to \mathbb{R}$ and $I_t : \mathcal{A} \to \mathbb{R}_{>0}$ compute the expected regret and expected information gain of actions with respect to the posterior distribution in round t.

The first variant of IDS, which is the main focus of Russo and Van Roy (2014a, 2018) and Kirschner et al. (2020a), chooses its action sampling distribution $\tilde{\pi}_t^{IDS}$ to satisfy,

$$\tilde{\pi}_t^{IDS} \in \operatorname*{argmin}_{\pi \in \mathcal{D}(\mathcal{A}_t)} \tilde{\Psi}_t^{IDS}(\pi), \text{ where } \tilde{\Psi}_t^{IDS}(\pi) = \frac{\tilde{\Delta}_t(\pi)^2}{\tilde{I}_t(\pi)}. \tag{14}$$

Here, $\mathcal{D}(\mathcal{A}_t)$ is a family of distributions over $\mathcal{A}_t$, and $\tilde{\Delta}_t$ and $\tilde{I}_t$ are analogs of the regret and information gain for distributions. Specifically, $\tilde{\Delta}_t(\pi) = \int_{a \in \mathcal{A}} \Delta_t(a) \mathrm{d}\pi(a)$, and $\tilde{I}_t(\pi) = \int_{a \in \mathcal{A}} I_t(a) \mathrm{d}\pi(a)$.

The second variant introduces a further tunable parameter $\lambda > 0$ and characterises the trade-off by a difference rather than a ratio, selecting its action sampling distribution π_t^{IDS} as follows,

$$\pi_t^{IDS} \in \operatorname*{argmin}_{\pi \in \mathcal{D}(\mathcal{A}_t)} \Psi_t^{IDS}(\pi), \text{ where } \Psi_t^{IDS}(\pi) = \Delta_t(\pi)^2 - \lambda I_t(\pi). \tag{15}$$

This second approach is mentioned in Russo and Van Roy (2014a, 2018), but has received less attention elsewhere. It can be shown to inherit the theoretical properties of the first variant if $\lambda \geq \tilde{\Psi}_t^{IDS}(\tilde{\pi}_t^{IDS})$ for every $t \in [T]$. We will show that this second approach is, conceptually, better suited to our problem. Since the first variant (i.e. the policy using (14)) has been more widely used we will refer to it as *traditional* IDS, and the second (i.e. the policy using (15)) as *tunable* IDS.

Notice that in the apple tasting problem, as there are only two actions in any round and as $A_t = 0$ has no associated information gain, we have $I_t(\pi_p) = pI_t(1)$, for any Bernoulli action selection distribution π_p where the probability of choosing class 1 is p. The optimisation problem (14) can be reduced to a line search over $p \in (0, 1)$, and the gradient of the objective can be written,

$$\frac{\mathrm{d}\tilde{\Psi}_t^{IDS}(\pi_p)}{\mathrm{d}p} = \frac{p^2 I_t(1) \left(\Delta_t(1) - \Delta_t(0)\right)^2 - I_t(1)\Delta_t(0)^2}{I_t(1)p}$$

It follows that $\tilde{\pi}_t^{IDS} = \pi_{\tilde{p}_t}$, where

$$\tilde{p}_t = \min\left(1, \frac{\Delta_t(0)}{|\Delta_t(1) - \Delta_t(0)|}\right).$$

Notice that $\tilde{p}_t$ is independent of $I_t(1)$ (so long as $I_t(1) > 0$), and secondly that for all $\Delta_t(1) \leq 2\Delta_t(0)$ traditional IDS chooses the label 1 with probability 1, even if $\Delta_t(1) > \Delta_t(0)$.

Remark 1 *We see that for traditional IDS, the* magnitude *of the information gain in a particular round is immaterial. As there is an action with no information, IDS prefers the information gaining action unless there is strong regret based reason to choose the no information action. Observe that* $\Delta_t(1)$ *must be three times as large as* $\Delta_t(0)$ *before traditional IDS would begin to prefer action 0. We argue that this property is not desirable in a contextual (or otherwise non-stationary) setting, where the information gain may change from round to round, and note that it occurs in more general settings.*

Considering the more general form of $\tilde{\Psi}_t^{IDS}$ *in* (14)*, it is clear that* $\tilde{\pi}_t^{IDS}$ *is invariant to any uniform scaling of the expected information function* $\tilde{I}_t$*. This is a potentially hazardous property in a range of situations, but seems to have the most pronounced effect in the setting where (potentially optimal) zero-information actions exist.*

In what follows we consider tunable IDS. Here the action selection distribution does depend on the magnitude of the information in a particular round, making it more suitable for our contextual problem. In particular, we have $\pi_t^{IDS} = \pi_{p_t}$, where

$$p_t = \max\left(0,\ \min\left(1,\ \frac{\lambda I_t(1)}{2(\Delta_t(1) - \Delta_t(0))^2} - \frac{\Delta_t(0)}{\Delta_t(1) - \Delta_t(0)}\right)\right). \tag{16}$$

4.1 Polya-Gamma Information Directed Sampling

As an alternative to our TS approach, we explore a tunable PG-IDS scheme, summarised in Algorithm 2.

The non-trivial difference between the PG-IDS and PG-TS schemes, from an implementation perspective, is the requirement to estimate posterior expectations rather than just draw a single sample from an approximate posterior. An exact IDS scheme would compute $\Delta_t(0)$, $\Delta_t(1)$, and $I_t(1)$ in each round according to,

$$\Delta_t(a) = \int_{\theta \in \mathbb{R}^d} \left[\mu(a, \theta) - \min_{a' \in \{0,1\}} \mu(a', \theta)\right] \mathrm{d}\pi_{t-1}(\theta), \quad a \in \{0, 1\},$$

and

$$I_t(1) = \mathbb{P}_{t-1}\left(C_t = 1\right) KL(\pi_{t-1}, \pi_{t-1} | C_t = 1) + \mathbb{P}_{t-1}\left(C_t = 0\right) KL(\pi_{t-1}, \pi_{t-1} | C_t = 0).$$

Such computations are of course infeasible due to the intractability of the posterior discussed in Section 3.

Instead, we rely on Monte Carlo estimators of these quantities. For expected regrets we have

$$\bar{\Delta}_t(a) = \frac{1}{M} \sum_{m=1}^{M} \left[\mu(a, \theta^{(m)}) - \min_{a' \in \{0,1\}} \mu(a', \theta^m)\right], \quad a \in \{0, 1\}. \tag{17}$$

Before defining the information gain, introduce the notation,

$$\ell(x, \theta, c) = \left(\sigma(x^\top \theta)\right)^c \left(1 - \sigma(x^\top \theta)\right)^{1-c}$$

as the likelihood contribution of the pair x, c for a given θ. To estimate $I_t(1)$ we require approximations of the normalising constants of the current posterior π_{t-1}, and the potential posteriors given the two possible updates. Define,

$$\bar{D}_t = \frac{1}{M} \sum_{m=1}^{M} \left[\prod_{s \in [t-1]: A_s = 1} \ell(x_s, \theta^{(m)}, C_s)\right], \text{ and}$$

$$\bar{D}_{t,c} = \frac{1}{M} \sum_{m=1}^{M} \left[\ell(x_t, \theta^{(m)}, c) \prod_{s \in [t-1]: A_s = 1} \ell(x_s, \theta^{(m)}, C_s)\right],$$

for $c \in \{0, 1\}$. These are used to estimate the KL divergences between the current posterior and the two potential posteriors in the next round. Define said estimates as,

$$\bar{K}_{t,c} = \frac{1}{M} \sum_{m=1}^{M} \log\left(\frac{\bar{D}_{t,c}}{\bar{D}_{t-1} \ell(x_t, \theta^{(m)}, c)}\right), \quad c \in \{0, 1\}.$$

Finally our estimate of the expected information gain follows as,

$$\bar{I}_t(1) = \bar{K}_{t,1} \sum_{m=1}^{M} \frac{\ell(x_t, \theta^{(m)}, 1)}{M} + \bar{K}_{t,0} \sum_{m=1}^{M} \frac{\ell(x_t, \theta^{(m)}, 0)}{M}. \tag{18}$$

We investigate the performance of PG-IDS alongside PG-TS in the next section.

Remark 2 *Both the PG-TS and PG-IDS approaches given in Algorithm 1 and 2 are perhaps the most straightforward possible in terms of their use of the Gibbs samples. It may be beneficial for instance to allow M to vary as a function of t, to remove some burn-in from the sample $\{\theta_t^{(1)}, \ldots, \theta_t^{(M)}\}$ before computing posterior expectations, in the case of PG-IDS, or to use separate samples to compute the regret estimates, and information gain. In the following section we have found strong empirical performance is achieved without such modifications, but it may be an interesting direction for future research to investigate whether these play a material role in the algorithms' performance.*

Inputs: Prior mean vector $\mathbf{b}$, Prior covariance matrix $\mathbf{B}$, Number of Gibbs iterations M, IDS-tuning parameter λ.

Initialise: $\mathcal{D} = \emptyset$, and $\theta_0^{(M)} \sim MVN_\Theta(\mathbf{b}, \mathbf{B})$.
for $t = 1, 2, \ldots$ **do**
  $\{\theta_t^{(1)}, \ldots, \theta_t^{(M)}\} \leftarrow \texttt{GIBBS}(\mathbf{b}, \mathbf{B}, M, \mathcal{D}, \theta_{t-1}^{(M)})$
  Receive context $\mathbf{x}_t \in \mathbb{R}^d$
  Compute regret estimates $\bar{\Delta}_t(0)$ and $\bar{\Delta}_t(1)$ according to (17)
  Compute information gain estimate $\bar{I}_t(1)$ according to (18)
  Compute IDS parameter p_t according to (16) using $\bar{\Delta}_t(0), \bar{\Delta}_t(1), \bar{I}_t(1)$
  Select action $a_t \sim Bern(p_t)$
  if $a_t = 1$ **do**
    Observe $l_t \in \{l_{01}, l_{11}\}$
    Augment $\mathcal{D} \leftarrow \mathcal{D} \cup \{x_t, \Phi_t(1)\}$
  end if
end for

Algorithm 2: PG-IDS

5 Simulations

We compare the PG-TS scheme in Algorithm 1 with a number of other algorithms. Firstly, the PG-IDS scheme given in Algorithm 2 and second, an ϵ-Greedy algorithm. The ϵ-Greedy approach chooses the action optimal with respect to the maximum likelihood estimate with probability $1-\epsilon$, otherwise it chooses randomly with equal probability. Finally, we consider the CBP-SIDE algorithm of Bartók and Szepesvári (2012), as investigated empirically in Lienert (2013). Pseudocode for the adaptation of this approach to the apple tasting problem is given in Appendix H. Fast

implementation of the PG-based algorithms is possible thanks to the 'BayesLogit' package in R (Polson et al., 2019).

For PG-IDS, and ϵ-Greedy we have selected the respective tunable parameters λ, and ϵ via a further empirical comparison outlined in Appendix I. The choices $\lambda = 0.05$, and $\epsilon = 0.1$ are argued to give the most robust performance among a range of possible values across various instances of our problem.

We consider three examples in terms of the true parameter, and context vector distribution, summarised below.

(i) Each dimension of the parameter θ^* is sampled uniformly from the interval $[-1, 1]$, with $d = 5$. Contexts are drawn from a zero-mean multivariate Gaussian with identity covariance matrix. We choose $l_{11} = 0.05$, $l_{01} = 0.4$, $l_{10} = 1$, and $T = 500$.

(ii) Each dimension of the parameter θ^* is sampled uniformly from the interval $[-1, 1]$ with probability 0.75, or fixed to 0 with probability 0.25. The number of dimensions is $d = 20$, and again contexts are drawn as in problem (i) but with common context variance of 8. We choose $l_{11} = 0.1$, $l_{01} = 0.7$, $l_{10} = 1$, and $T = 1000$.

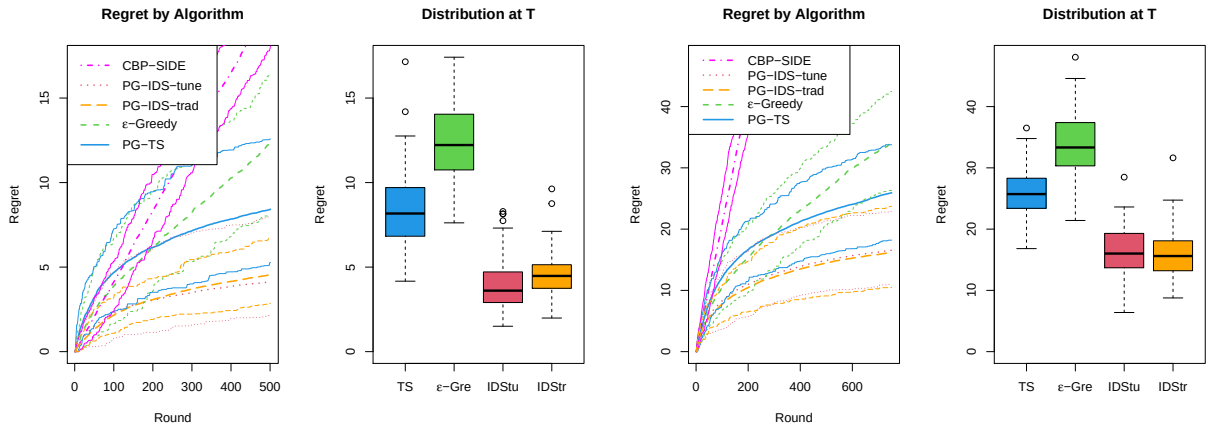
(iii) $\theta^* = 1$ and contexts are sampled from a Gaussian distribution with standard deviation 0.025. The mean linearly increases from $\mu = -0.1$ to $\mu = 0$ over the $T = 500$ rounds. We choose $l_{11} = 0$ and $l_{01} = l_{10} = 1$.

Problems (i) and (ii) represent typical apple tasting scenarios, and allow the comparison of our various algorithms. Problem (iii) gives a simple, uncluttered example of a scenario where the traditional variant of IDS performs poorly.

Figure 1 shows the performance of the algorithms on the problems described above. For problems (i) and (ii) we see similar behaviour, the two PG-IDS algorithms perform similarly and incur a lower average regret than the other approaches. PG-TS is successful in learning to select optimal actions but does so more slowly, and CBP-SIDE and ϵ-Greedy incur a regret which appears to grow linearly throughout the experiment. ϵ-Greedy could perhaps be improved by allowing ϵ to decay as a function of t, but this represents yet another possibly problem-dependent tuning step. As for CBP-SIDE, the observed behaviour is a result of its overly conservative approach. This is confirmed below by considering the precision and recall of the algorithms.

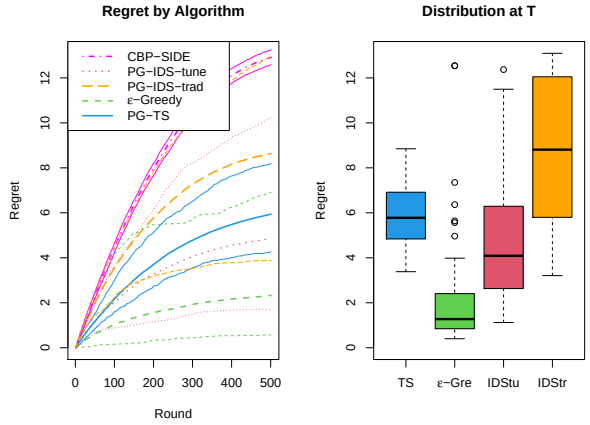
In problem (iii) where every context is sampled close to the classification boundary, with a strong tendency towards class 0, a different behaviour is observed. Besides PG-TS and CBP-SIDE, each of the algorithms displays a more variable performance, and the traditional variant of PG-IDS incurs noticeably larger regret than PG-TS and the tunable variant of PG-IDS. The mean regret of PG-TS and the tunable variant of PG-IDS are very similar. While ϵ-Greedy sometimes outperforms the more principled approaches, due to fixing an accurate initial estimate of θ^*, this overly exploitative behaviour sometimes fails and leads to highly skewed distribution of regret.

The phenomenon of a greedy approach sometimes outperforming more complex attempts to balance exploration and exploitation in contextual problems is not unprecedented. Indeed, in the contextual bandit literature, a number of recent works (Bastani et al., 2017; Kannan et al., 2018; Raghavan et al., 2020; Jedor et al., 2021) have observed that if the contexts arising are sufficiently variable, a greedy decision-maker can gain appropriate information without explicitly attempting to explore in favour of exploiting.



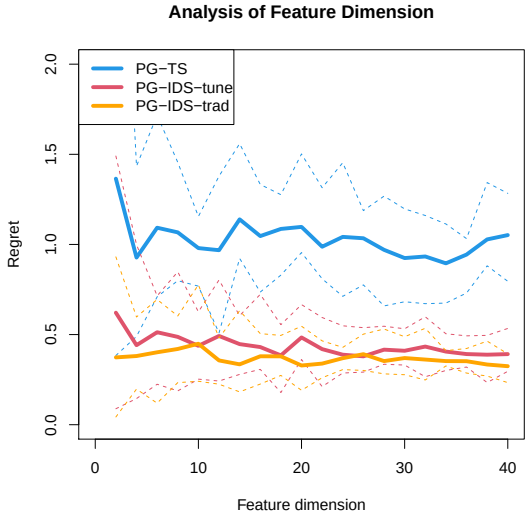
(a) Problem (i)

(b) Problem (ii)

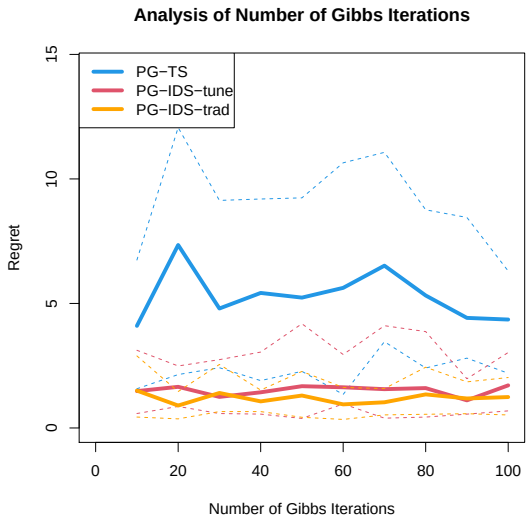


(c) Problem (iii)

Figure 1: Regret of algorithms on Problems (i), (ii) and (iii), over 50 replications. The green lines denote the ϵ-Greedy policy with $\epsilon = 0.1$, the red lines denote the PG-IDS policy with $\lambda = 0.2$, the magenta lines denote the CBP-SIDE policy, and the blue lines denote the PG-TS policy. In each case 90% empirical confidence regions are plotted around the median trajectory. The boxplots in the right-hand panel show the distribution of the final regret at time T.



(a) Regret of algorithms scaled by $\sqrt{d}$, plotted as d varies.



(b) Regret of algorithms plotted as M varies.

Figure 2: Plots showing the effects of problem and algorithm parameters on regret.

The regret is presumed to be the true measure of interest in the apple tasting problem, and to appropriately weight the impact of false positives and false negatives. Nevertheless it is informative to see in what proportion the algorithms make the two classes of error. To this end, in Tables 1, 2, and 3 we report the *precision* and *recall* of the apple tasting algorithms. The precision is the proportion of true class 1 examples among all those labelled class 1 by the algorithm, the recall is the proportion of class 1 examples correctly labelled as class 1 by the algorithm.

We observe that CBP-SIDE always has a recall of 1.00, and while we would expect this to reduce in problems with a longer horizon, this is indicative of the fact that CBP-SIDE's confidence region for θ^* is so wide that its optimism leads to using action 1 in almost every instance. As such it correctly labels all true class 1 instances, but misclassifies a large (relative to the other approaches) proportion of class 0 instances as being class 1.

In problems (i) and (ii) most of the other approaches have well balanced precision and recall, suggesting that misclassifications of both types occur with roughly similar frequency. In problem (iii), however, the traditional variant of PG-IDS shows a similar behaviour to CBP-SIDE, in that its recall is much larger than the algorithms with smaller average regret. This is likely because it suffers the issue postulated in Section 4. Specifically, that because of the construction of the traditional IDS sampling distribution, it continues to favour the information gaining action even when the regret of the no-information action is smaller, and the information gained per observation is minimal.

Finally, we have also investigated the effect of the dimension d of the unknown parameter, and of the number of Gibbs iterations M, on the regret of the algorithms. Figure 2 displays the results of these experiments. In the setting of Figure 2(a), the dimension is varied between $d = 2$ and $d = 40$, with the elements of θ^* being drawn uniformly at random from the interval $[-1, 1]$ and M fixed at 15. In the setting of Figure 2(b), we fixed $d = 5$, again drew the elements of θ^*

Algorithm	Precision	Recall
PG-TS	0.93	0.90
PG-IDS-tune	0.94	0.94
PG-IDS-trad	0.91	0.96
Greedy	0.93	0.91
CBP-SIDE	0.74	1.00

Table 1: Precision and Recall on Problem (i)

Algorithm	Precision	Recall
PG-TS	0.89	0.88
PG-IDS-tune	0.91	0.92
PG-IDS-trad	0.88	0.95
Greedy	0.88	0.87
CBP-SIDE	0.53	1.00

Table 2: Precision and Recall on Problem (ii)

Algorithm	Precision	Recall
PG-TS	0.13	0.56
PG-IDS-tune	0.26	0.75
PG-IDS-trad	0.17	0.92
Greedy	0.49	0.85
CBP-SIDE	0.10	1.00

Table 3: Precision and Recall on Problem (iii)

uniformly from $[-1, 1]$, and varied the number of Gibbs iterations per time-step between $M = 10$ and $M = 100$.

In Figure 2(a) we plot the mean regret for each investigated dimension d and algorithm with a 90% empirical confidence interval, all scaled by $\sqrt{d}$ to investigate the relationship between regret and dimension. We see a mostly constant relationship, within the tolerance of statistical noise, which seems to validate the theoretical result on regret. In Figure 2(b) we see again see a fairly constant relationship, suggesting that the performance of the algorithms is fairly robust to the number of Gibbs iterations. This is an encouraging finding as it suggests a more computationally burdensome choice of the parameter M is not necessary for strong performance.

6 Conclusion

In this paper we have explored the use of heuristic Bayesian decision-making rules for logistic contextual apple tasting problems. We have shown that both Thompson Sampling and Information Directed Sampling methods are highly efficient for such problems, and indeed more so than confidence bound based approach of Bartók and Szepesvári (2012). We have also established a theoretical justification of the strong performance of Thompson Sampling through a bound on its Bayesian regret, and of Pólya-Gamma Thompson Sampling by considering its asymptotic behaviour. The extension of these results to Information Directed Sampling will be significantly more complex due to the choice of actions with respect to posterior expectations rather than individual samples[6]. Other further research may explore the application of Bayesian approaches to more complex contextual partial monitoring problems, incorporating the insights of Lattimore and Szepesvári (2019), or to other variants of apple tasting, such as the fairness-enforcing variant considered by Bechavod et al. (2019), or batched setting of Jiang et al. (2021).

[6]Note that this difficulty is specific to the present setting where Θ is not a finite set. The challenge arises in defining a suitable analog of the compressed random variable $\tilde{\theta}_t^*$ as in the proof of Theorem 1.

Acknowledgements: The authors gratefully acknowledge the support of the Next Generation Converged Digital Infrastructure project (EP/R004935/1) funded by the EPSRC and BT, and thank Fabrizio Leisen, Christopher Nemeth, Ciara Pike-Burke, and Christopher Sherlock for helpful conversations during the preparation of this manuscript.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320.

Agarwal, A. (2013). Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228.

Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107.

Agrawal, S. and Goyal, N. (2013b). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, 2(4):319–342.

Antos, A., Bartók, G., Pál, D., and Szepesvári, C. (2013). Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99.

Arumugam, D. and Van Roy, B. (2020). Randomized value functions via posterior state-abstraction sampling. *arXiv preprint arXiv:2010.02383*.

Bartók, G., Foster, D. P., Pál, D., Rakhlin, A., and Szepesvári, C. (2014). Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997.

Bartók, G., Pál, D., and Szepesvári, C. (2010). Toward a classification of finite partial-monitoring games. In *International Conference on Algorithmic Learning Theory*, pages 224–238. Springer.

Bartók, G., Pál, D., and Szepesvári, C. (2011). Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 133–154.

Bartók, G. and Szepesvári, C. (2012). Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pages 305–319. Springer.

Bastani, H., Bayati, M., and Khosravi, K. (2017). Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. (2019). Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, pages 8974–8984.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2011). Learning noisy linear classifiers via adaptive and selective sampling. *Machine learning*, 83(1):71–102.

Cesa-Bianchi, N., Gentile, C., and Orabona, F. (2009). Robust bounds for classification via selective sampling. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 121–128.

Cesa-Bianchi, N., Long, P. M., and Warmuth, M. K. (1996). Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Choi, H. M. and Hobert, J. P. (2013). The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064.

Chow, C.-K. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Dekel, O., Gentile, C., and Sridharan, K. (2012). Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*.

Dong, S., Ma, T., and Van Roy, B. (2019). On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160.

Dong, S. and Van Roy, B. (2018). An information-theoretic analysis for thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165.

Dumitrascu, B., Feng, K., and Engelhardt, B. (2018). Pg-ts: Improved thompson sampling for logistic contextual bandits. In *Advances in Neural Information Processing Systems*, pages 4624–4633.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Gentile, C. and Orabona, F. (2014). On multilabel classification and ranking with bandit feedback. *The Journal of Machine Learning Research*, 15(1):2451–2487.

Ghosal, S., Ghosh, J. K., Samanta, T., et al. (1995). On convergence of posterior distributions. *The Annals of Statistics*, 23(6):2145–2152.

Grant, J. A., Boukouvalas, A., Griffiths, R.-R., Leslie, D. S., Vakili, S., and De Cote, E. M. (2019). Adaptive sensor placement for continuous spaces. In *International Conference on Machine Learning*, pages 2385–2393.

Grant, J. A. and Leslie, D. S. (2020). On thompson sampling for smoother-than-lipschitz bandits. In *23rd International Conference on Artificial Intelligence and Statistics*.

Helmbold, D. P., Littlestone, N., and Long, P. M. (1992). Apple tasting and nearly one-sided learning. In *Proceedings., 33rd Annual Symposium on Foundations of Computer Science*, pages 493–502. IEEE Computer Society.

Helmbold, D. P., Littlestone, N., and Long, P. M. (2000). Apple tasting. *Information and Computation*, 161(2):85–139.

Jedor, M., Louëdec, J., and Perchet, V. (2021). Be greedy in multi-armed bandits. *arXiv preprint arXiv:2101.01086*.

Jiang, H., Jiang, Q., and Pacchiano, A. (2021). Learning the truth from only one side of the story. In *International Conference on Artificial Intelligence and Statistics*, pages 2413–2421. PMLR.

Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, pages 2227–2236.

Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384.

Kirschner, J., Lattimore, T., and Krause, A. (2020a). Information directed sampling for linear partial monitoring. In *Conference on Learning Theory*, pages 2328–2369. PMLR.

Kirschner, J., Lattimore, T., Vernade, C., and Szepesvári, C. (2020b). Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*.

Lattimore, T. and Szepesvári, C. (2019). An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

Lienert, I. (2013). Exploiting side information in partial monitoring games: An empirical study of the cbp-side algorithm with applications to procurement. Master's thesis, Eidgenössische Technische Hochschule Zürich, Department of Computer Science.

Littlestone, N. (1990). Mistake bounds and logarithmic linear-threshold learning algorithms.

Liu, D., Zhang, P., and Zheng, Q. (2015). An efficient online active learning algorithm for binary classification. *Pattern Recognition Letters*, 68:22–26.

Liu, F., Buccapatnam, S., and Shroff, N. (2018). Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lorentz, G. (1966). Metric entropy and approximation. *Bulletin of the American Mathematical Society*, 72(6):903–937.

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

Mazumdar, E., Pacchiano, A., Ma, Y.-a., Bartlett, P. L., and Jordan, M. I. (2020). On thompson sampling with langevin algorithms.

Monteleoni, C. and Kaariainen, M. (2007). Practical online active learning for classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Orabona, F. and Cesa-Bianchi, N. (2011). Better algorithms for selective sampling. In *International Conference on Machine Learning*, pages 433–440. Omnipress.

Phan, M., Yadkori, Y. A., and Domke, J. (2019). Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, pages 8804–8813.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Polson, N. G., Scott, J. G., and Windle, J. (2019). Package 'bayeslogit'.

Raghavan, M., Slivkins, A., Vaughan, J. W., and Wu, Z. S. (2020). Greedy algorithm almost dominates in smoothed contextual bandits. *arXiv preprint arXiv:2005.10624*.

Russo, D. and Van Roy, B. (2014a). Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591.

Russo, D. and Van Roy, B. (2014b). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Russo, D. and Van Roy, B. (2018). Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sayedi, A., Zadimoghaddam, M., and Blum, A. (2010). Trading off mistakes and don't-know predictions. In *Advances in Neural Information Processing Systems*, pages 2092–2100.

Sculley, D. (2007). Practical learning from one-sided feedback. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 609–618.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Tsuchiya, T., Honda, J., and Sugiyama, M. (2020). Analysis and design of thompson sampling for stochastic partial monitoring. *arXiv preprint arXiv:2006.09668*.

Urteaga, I. and Wiggins, C. (2018). Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 698–706. PMLR.

Wiener, Y. and El-Yaniv, R. (2011). Agnostic selective classification. In *Advances in Neural Information Processing Systems*, pages 1665–1673.

Windle, J., Polson, N. G., and Scott, J. G. (2014). Sampling pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*.

Yang, Y. and Dunson, D. B. (2013). Sequential markov chain monte carlo. *arXiv preprint arXiv:1308.3861*.

Yang, Y. and Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121.

Ying, Y. and Zhou, D.-X. (2006). Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788.

A Proof of Proposition 1

The first step is a functional version of Lemma 1 of Dong and Van Roy (2018) for non-finite Θ, which requires a different proof technique. The proof of Lemma 1 is given in Appendix B.

Lemma 1 *Let Λ be a compact space in $\mathbb{R}^d$, and f and g be two functions from Λ to $\mathbb{R}$. Let π be a density function on Λ such that $\pi(\lambda) > 0$ for all $\lambda \in \Lambda$. There exist $\lambda, \lambda' \in \Lambda$ (possibly $\lambda = \lambda'$) and $p \in (0,1)$ such that*

$$\begin{aligned} pf(\lambda) + (1-p)f(\lambda') &\leq \mathbb{E}_\pi(f), \text{ and} \\ pg(\lambda) + (1-p)g(\lambda') &\leq \mathbb{E}_\pi(g), \end{aligned}$$

where $\mathbb{E}_\pi(h) = \int_{\lambda \in \Lambda} h(\lambda)\pi(\mathrm{d}\lambda)$ for a function $h : \Lambda \to \mathbb{R}$.

Applying Lemma 1 on Θ_k, the k^{th} element of the partition, with the functions $f(\cdot) = -\mathbb{E}_{t-1}(\mu_t(\alpha_t(\cdot), \theta^*))$ and $g(\cdot) = I_{t-1}(\phi_\epsilon; \Phi_t(\alpha_t(\cdot)))$ and the density $\pi(\theta_t | \theta_t \in \Theta_k)$ we have that for each $k \in [K]$ and $t \in [T]$ there exists two parameters $\theta_1^{k,t}, \theta_2^{k,t} \in \Theta_k$ and a constant $r_{k,t} \in (0,1)$ such that,

$$\begin{aligned} r_{k,t}\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t\left(\theta_1^{k,t}\right), \theta^*\right)\right) + (1-r_{k,t})\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t\left(\theta_2^{k,t}\right), \theta^*\right)\right) \\ \geq \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\theta_t), \theta^*\right) \mid \theta_t \in \Theta_k\right), \end{aligned} \tag{19}$$

and

$$r_{k,t}I_{t-1}\left(\phi_\epsilon; \Phi_t\left(\alpha_t\left(\theta_1^{k,t}\right)\right)\right) + (1-r_{k,t})I_{t-1}\left(\phi_\epsilon; \Phi_t\left(\alpha_t\left(\theta_2^{k,t}\right)\right)\right)$$

$$\leq I_{t-1}\left(\phi_\epsilon; \Phi_t(\alpha_t(\theta_t)) \mid \theta_t \in \Theta_k\right). \tag{20}$$

Let $\tilde{\theta}_t^*$ be a random variable with the following conditional distribution given ϕ_ϵ,

$$P\left(\tilde{\theta}_t^* = \theta_1^{k,t} \mid \phi_\epsilon = k\right) = r_{k,t},\ P\left(\tilde{\theta}_t^* = \theta_2^{k,t} \mid \phi_\epsilon = k\right) = 1 - r_{k,t}, \quad \forall k \in [K]. \tag{21}$$

Immediately, we have that $\tilde{\theta}_t^*$ satisfies property (i) as is independent of θ^* given ϕ_ϵ. Using (19) and (20), properties (ii) and (iii) can be shown to hold for $\tilde{\theta}_t^*$.

To show (ii), we first define, for every $k \in [K]$ and $t \in [T]$,

$$\begin{aligned} D_{k,t} = r_{k,t}\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t\left(\theta_1^{k,t}\right), \theta^*\right)\right) + (1 - r_{k,t})\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t\left(\theta_2^{k,t}\right), \theta^*\right)\right) \\ - \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\theta_t), \theta^*\right) \mid \theta_t \in \Theta_k\right). \end{aligned}$$

By (19), each $D_{k,t} \geq 0$. Thus, we also have,

$$\mathbb{E}_{t-1}\left(\mu_t(\alpha_t(\tilde{\theta}_t), \theta^*) - \mu_t(\alpha(\theta_t), \theta^*)\right) = \sum_{k=1}^{K} P(\theta_t \in \Theta_k) D_{k,t} \geq 0.$$

Property (ii) then follows using this result in the first inequality,

$$\begin{aligned} &\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\theta_t), \theta^*\right) - \mu_t\left(\alpha_t(\theta^*), \theta^*\right)\right) - \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right) \\ &= \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) - \mu_t\left(\alpha_t(\theta^*), \theta^*\right)\right) - \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\theta_t), \theta^*\right)\right) \\ &\leq \mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) - \mu_t\left(\alpha_t(\theta^*), \theta^*\right)\right) \leq \epsilon. \end{aligned}$$

Here the final inequality holds since $\tilde{\theta}_t^*$ and θ^* are always in the same cell of the partition, and (ii) follows by simple rearrangement.

Considering the information gain, we show property (iii) as follows, where the second and final equalities use the independence of θ_t and $\tilde{\theta}_t$ from ϕ_ϵ conditioned on $\mathcal{H}_{t-1}$, and the inequality uses (20),

$$\begin{aligned} I_{t-1}\left(\phi_\epsilon; \left(\tilde{\theta}_t, \Phi_t(\alpha_t(\tilde{\theta}_t))\right)\right) &= I_{t-1}\left(\phi_\epsilon; \tilde{\theta}_t\right) + I_{t-1}\left(\phi_\epsilon; \left(\Phi_t(\alpha_t(\tilde{\theta}_t) \mid \tilde{\theta}_t\right)\right) \\ &= I_{t-1}\left(\phi_\epsilon; \left(\Phi_t(\alpha_t(\tilde{\theta}_t) \mid \tilde{\theta}_t\right)\right) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{2} P\left(\theta_t \in \Theta_k\right) P\left(\tilde{\theta}_t = \theta_i^{k,t} \mid \theta_t \in \Theta_k\right) I_{t-1}\left(\phi_\epsilon; \Phi_t(\alpha_t(\tilde{\theta}_i^{k,t}))\right) \\ &\leq \sum_{k=1}^{K} P\left(\theta_t \in \Theta_k\right) I_{t-1}\left(\phi_\epsilon; \Phi_t(\alpha_t(\theta_t)) \mid \theta_t \in \Theta_k\right) \\ &= I_{t-1}\left(\phi_\epsilon; \Phi_t(\alpha_t(\theta_t))\right) = I_{t-1}\left(\phi_\epsilon; (\theta_t, \Phi_t(\alpha_t(\theta_t)))\right). \square \end{aligned}$$

B Proof of Lemma 1

To simplify notation, we first assume without loss of generality (as the following can be achieved by shifting f and g) that

$$\int_\Lambda \pi(\mathrm{d}\lambda) f(\lambda) = \int_\Lambda \pi(\mathrm{d}\lambda) g(\lambda) = 0$$

Then we define $C_f(\lambda, \lambda'; p) = pf(\lambda) + (1-p)f(\lambda')$ and similarly $C_g(\lambda, \lambda'; p) = pg(\lambda) + (1-p)g(\lambda')$. To prove Lemma 1, we wish to show that $\exists p \in [0,1], \lambda, \lambda' \in \Lambda$ such that

$$\max\left[C_f(\lambda, \lambda'; p), C_g(\lambda, \lambda'; p)\right] \leq 0. \tag{22}$$

It will also be helpful to define the following subsets of Λ,

$$\Lambda_f = \{\lambda \in \Lambda \,:\, f(\lambda) \leq 0\},$$
$$\Lambda_g = \{\lambda \in \Lambda \,:\, g(\lambda) \leq 0\}.$$

We note that if $\Lambda_f \cap \Lambda_g \neq \emptyset$, i.e. there exists $\lambda \in \Lambda$ such that $f(\lambda) \leq 0$ and $g(\lambda) \leq 0$, then the proof of Lemma 1 is trivial. Choosing any $\lambda^* \in \Lambda_f \cap \Lambda_g$, we have $C_f(\lambda^*, \lambda^*; p) \leq 0$ and $C_g(\lambda^*, \lambda^*; p) \leq 0$ for all $p \in (0,1)$. The more challenging case is where $\Lambda_f \cap \Lambda_g = \emptyset$, i.e. there exists no $\lambda \in \Lambda$ such that $f(\lambda) \leq 0$ and $g(\lambda) \leq 0$, which we will consider in the remainder of the proof. In this case, for all $\lambda \in \Lambda_f$, we have $f(\lambda) \leq 0 < g(\lambda)$ and for all $\lambda' \in \Lambda_g$, we have $g(\lambda') \leq 0 < f(\lambda')$.

Since both C functions are linear in p, it follows that $\min_p \max\{C_f(\lambda, \lambda'; p), C_g(\lambda, \lambda'; p)\}$ is achieved where the two C targets are equal (C_f increases from $f(\lambda) < 0$ to $f(\lambda') > 0$ and C_g decreases, so the max decreases to the intersection then increases again). Hence the minimising p is

$$p_{min}(\lambda, \lambda') = \frac{g(\lambda') - f(\lambda')}{f(\lambda) - f(\lambda') - g(\lambda) + g(\lambda')}$$

and the minimum takes value

$$\frac{f(\lambda)g(\lambda') - f(\lambda')g(\lambda)}{f(\lambda) - f(\lambda') - g(\lambda) + g(\lambda')}.$$

Now, $f(\lambda) \leq 0$, $g(\lambda) > 0$, $f(\lambda') > 0$ and $g(\lambda') \leq 0$ so the denominator is strictly negative. Thus (22) is satisfied when there exist $\lambda \in \Lambda_f$ and $\lambda' \in \Lambda_g$ such that

$$f(\lambda)g(\lambda') - f(\lambda')g(\lambda) \geq 0 \tag{23}$$

We show that there exist such λ, λ' using a proof by contradiction. We assume the converse of (23):

$$\frac{f(\lambda)g(\lambda')}{f(\lambda')g(\lambda)} < 1, \quad \forall \lambda \in \Lambda_f,\ \lambda' \in \Lambda_g. \tag{24}$$

Let $D_f = \max_{\lambda \in \Lambda_f} g(\lambda)/f(\lambda)$ and $D_g = \max_{\lambda' \in \Lambda_g} f(\lambda')/g(\lambda')$. Both quantities exist and are strictly negative, since Λ_f and Λ_g are compact. Furthermore (24) implies that $f(\lambda) > D_g g(\lambda)$ for all $\lambda \in \Lambda_f$, that $g(\lambda') > D_f f(\lambda')$ for all $\lambda' \in \Lambda_g$, and that $D_f D_g < 1$. It follows immediately that $E(f(\lambda) \,|\, \lambda \in \Lambda_f) \geq D_g E(g(\lambda) \,|\, \lambda \in \Lambda_f)$ and $E(g(\lambda') \,|\, \lambda' \in \Lambda_g) \geq D_f E(f(\lambda') \,|\, \lambda' \in \Lambda_g)$. Hence (22) implies that

$$\frac{E(f(\lambda)|\lambda \in \Lambda_f)}{E(g(\lambda)|\lambda \in \Lambda_f)} \frac{E(g(\lambda')|\lambda' \in \Lambda_g)}{E(f(\lambda')|\lambda' \in \Lambda_g)} \leq D_f D_g < 1. \tag{25}$$

We will however show that the opposite must hold. Notice that $\int_{\Lambda_f} \pi(\mathrm{d}\lambda)(-f(\lambda)) \geq \int_{\Lambda_g} \pi(\mathrm{d}\lambda') f(\lambda')$, and $\int_{\Lambda_g} \pi(\mathrm{d}\lambda')(-g(\lambda')) \geq \int_{\Lambda_f} \pi(\mathrm{d}\lambda) g(\lambda)$ since $\Lambda_f \cap \Lambda_g = \emptyset$ and the expectation of both f and g is 0. It follows from these properties that

$$\frac{E(f(\lambda) \mid \lambda \in \Lambda_f) E(g(\lambda') \mid \lambda' \in \Lambda_g)}{E(f(\lambda') \mid \lambda \in \Lambda_g) E(g(\lambda) \mid \lambda \in \Lambda_f)} = \frac{\int_{\Lambda_f} -\frac{\pi(\mathrm{d}\lambda)}{\int_{\Lambda_f} \pi(d\theta)} f(\lambda) \cdot \int_{\Lambda_g} -\frac{\pi(\mathrm{d}\lambda')}{\int_{\Lambda_g} \pi(d\theta)} g(\lambda')}{\int_{\Lambda_g} \frac{\pi(\mathrm{d}\lambda')}{\int_{\Lambda_g} \pi(d\theta)} f(\lambda') \cdot \int_{\Lambda_f} \frac{\pi(\mathrm{d}\lambda)}{\int_{\Lambda_f} \pi(d\theta)} g(\lambda)}$$

$$= \frac{\int_{\Lambda_f} -\pi(\mathrm{d}\lambda) f(\lambda) \cdot \int_{\Lambda_g} -\pi(\mathrm{d}\lambda') g(\lambda')}{\int_{\Lambda_g} \pi(\mathrm{d}\lambda') f(\lambda') \cdot \int_{\Lambda_f} \pi(\mathrm{d}\lambda) g(\lambda)}$$
$$\geq 1 \tag{26}$$

Equation (26) is a direct contradiction of (24) and it follows that there must exist $\lambda \in \Lambda_f$ and $\lambda' \in \Lambda_g$ such that (23) holds, and Lemma 1 is proven. □

C Proof of Proposition 2

Recall the definition of the information ratio of the random variables $\tilde{\theta}_t^*$ and $\tilde{\theta}_t$ as

$$\Gamma_t\left(\tilde{\theta}_t^*, \tilde{\theta}_t\right) = \frac{\left[\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right)\right]^2}{I_{t-1}\left(\tilde{\theta}_t^*; (\tilde{\theta}_t, \Phi_t(\alpha_t(\tilde{\theta}_t)))\right)},$$

where the expectation in the numerator is taken over the random variables $\tilde{\theta}_t^*, \tilde{\theta}_t$ and θ^* conditional on the history $\mathcal{H}_{t-1}$ (recall that $\tilde{\theta}_t^*$ is the compressed version of θ^*, whereas $\tilde{\theta}_t$ has the same marginal distribution as $\tilde{\theta}_t^*$ but is conditionally independent of θ^* and $\tilde{\theta}_t^*$ given $\mathcal{H}_{t-1}$). It is worth noting that although the information gain in a particular round *can* take value zero, the quantity in the denominator is an expectation of the mutual information over the distribution of $\alpha_t(\tilde{\theta}_t)$ and the signal $\Phi_t(\alpha_t(\tilde{\theta}_t))$ and as such will be non-zero so long as $\pi_{t-1}(\Theta_t) > 0$.

To bound Γ_t, we first rewrite the root of the numerator in the information ratio. Denote by $\tilde{\pi}_t(\tilde{\theta}_t)$ and $\tilde{\pi}_t^*(\tilde{\theta}_t^*)$ the marginal probability mass functions of $\tilde{\theta}_t$ and $\tilde{\theta}_t^*$ respectively conditional on $\mathcal{H}_{t-1}$ and denote by $\pi_t^*(\theta^*)$ and $\pi_t^*(\theta^* \,|\, \tilde{\theta}_t^*)$ the marginal and conditional-on-θ_t^* density functions of θ^* given $\mathcal{H}_{t-1}$, noting that these random variables are constructed such that $\tilde{\pi}_t(\theta) = \tilde{\pi}_t^*(\theta)$ for all θ and, conditional on $\mathcal{H}_t$, $\tilde{\theta}_t$ is independent of both $\tilde{\theta}_t^*$ and θ^*. We find that

$$\begin{aligned}
&\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right) \\
&= \sum_{\tilde{\theta}_t} \sum_{\tilde{\theta}_t^*} \int \left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right) \tilde{\pi}_t(\tilde{\theta}_t) \tilde{\pi}_t^*(\tilde{\theta}_t^*) \pi_t^*(\theta^* \,|\, \tilde{\theta}_t^*)\, \mathrm{d}\theta^* \\
&= \sum_{\tilde{\theta}_t} \int \mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) \tilde{\pi}_t(\tilde{\theta}_t) \pi_t^*(\theta^*)\, \mathrm{d}\theta^* - \sum_{\tilde{\theta}_t^*} \int \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) \tilde{\pi}_t^*(\tilde{\theta}_t^*) \pi_t^*(\theta^* \,|\, \tilde{\theta}_t^*)\, \mathrm{d}\theta^* \\
&= \sum_{\tilde{\theta}_t^*} \left\{ \int \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) \pi_t^*(\theta^*)\, \mathrm{d}\theta^* - \int \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) \pi_t^*(\theta^* \,|\, \tilde{\theta}_t^*)\, \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*). \\
&= \sum_{\tilde{\theta}_t^*} \left\{ \int \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right) \left(\pi_t^*(\theta^*) - \pi_t^*(\theta^* \,|\, \tilde{\theta}_t^*)\right)\, \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*).
\end{aligned}$$

Recall from the definition of Θ_t that when $\tilde{\theta}_t^* \in \Theta_t$, the optimal response is $\alpha_t(\tilde{\theta}_t^*) = 1$, and (from (1)) the loss function is $1 - (l_{11} - 1)\sigma(x_t^T \theta^*)$. When $\tilde{\theta}_t^* \in \Theta \setminus \Theta_t$ the optimal action is $\alpha_t(\tilde{\theta}_t^*) = 0$, and the loss function is $l_{01}\sigma(x_t^T \theta_t^*)$. Continuing the above derivation, and noting that $\int \pi_t^*(\theta^*)\mathrm{d}\theta^* = \int \pi_t^*(\theta^* | \tilde{\theta}_t^*)\mathrm{d}\theta^* = 1$), we have,

$$\mathbb{E}_{t-1}\left(\mu_t\left(\alpha_t(\tilde{\theta}_t), \theta^*\right) - \mu_t\left(\alpha_t(\tilde{\theta}_t^*), \theta^*\right)\right)$$

$$
\begin{aligned}
&= \sum_{\tilde{\theta}_t^* \in \Theta_t} \left\{ \int \left(1 + (l_{11} - 1)\sigma(x_t^\top \theta^*)\right) \left(\pi_t^*(\theta^*) - \pi_t^*(\theta^* \mid \tilde{\theta}_t^*)\right) \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&\quad + \sum_{\tilde{\theta}_t^* \in \Theta \setminus \Theta_t} \left\{ \int l_{01}\sigma(x_t^\top \theta^*) \left(\pi_t^*(\theta^*) - \pi_t^*(\theta^* \mid \tilde{\theta}_t^*)\right) \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&= \sum_{\tilde{\theta}_t^* \in \Theta_t} \left\{ \int (l_{11} - 1)\sigma(x_t^\top \theta^*) \left(\pi_t^*(\theta^*) - \pi_t^*(\theta^* \mid \tilde{\theta}_t^*)\right) \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&\quad + \sum_{\tilde{\theta}_t^* \in \Theta \setminus \Theta_t} \left\{ \int l_{01}\sigma(x_t^\top \theta^*) \left(\pi_t^*(\theta^*) - \pi_t^*(\theta^* \mid \tilde{\theta}_t^*)\right) \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&\leq \max(l_{01}, 1 - l_{11}) \sum_{\tilde{\theta}_t^*} \left| \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^*) \mathrm{d}\theta^* - \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^* \mid \tilde{\theta}_t^*) \mathrm{d}\theta^* \right| \tilde{\pi}_t^*(\tilde{\theta}_t^*). \qquad (27)
\end{aligned}
$$

Next, consider the denominator. We have

$$
\begin{aligned}
I_{t-1}\left(\tilde{\theta}_t^*;\ (\tilde{\theta}_t, \Phi_t(\alpha_t(\tilde{\theta}_t)))\right) &= I_{t-1}\left(\tilde{\theta}_t^*;\ \tilde{\theta}_t\right) + I_{t-1}\left(\tilde{\theta}_t^*;\ \Phi_t(\alpha_t(\tilde{\theta}_t)) \mid \tilde{\theta}_t\right) \\
&= I_{t-1}\left(\tilde{\theta}_t^*;\ \Phi_t(\alpha_t(\tilde{\theta}_t)) \mid \tilde{\theta}_t\right) \\
&= \sum_{\theta} I_{t-1}\left(\tilde{\theta}_t^*;\ \Phi_t(\alpha_t(\theta))\right) \tilde{\pi}_t(\theta) \\
&= \tilde{\pi}_t(\Theta_t) I_{t-1}\left(\tilde{\theta}_t^*;\ \Phi_t(1)\right),
\end{aligned}
$$

where we have used the conditional independence of $\tilde{\theta}_t$ and $\tilde{\theta}_t^*$ and the fact that the information gain is zero if $A_t = 0$ is chosen. Then, rewriting the mutual information in terms of KL-divergence and applying Pinsker's inequality, we have the following lower bound,

$$
\begin{aligned}
&\tilde{\pi}_t(\Theta_t) I_{t-1}\left(\tilde{\theta}_t^*;\ \Phi_t(1)\right) \\
&= \tilde{\pi}_t(\Theta_t) \sum_{\tilde{\theta}_t^*} KL\left[\pi_{t-1}\left(\Phi_t(1) \mid \tilde{\theta}_t^*\right) \ ||\ \pi_{t-1}\left(\Phi_t(1)\right)\right] \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&\geq 2\tilde{\pi}_t(\Theta_t) \sum_{\tilde{\theta}_t^*} d_{TV}\left(\pi_{t-1}\left(\Phi_t(1) \mid \tilde{\theta}_t^*\right),\ \pi_{t-1}\left(\Phi_t(1)\right)\right)^2 \tilde{\pi}_t^*(\tilde{\theta}_t^*) \\
&= 2\tilde{\pi}_t(\Theta_t) \sum_{\tilde{\theta}_t^*} \left\{ \int \sigma\left(x_t^\top \theta^*\right) \pi_t^*(\theta^*) \mathrm{d}\theta^* - \int \sigma\left(x_t^\top \theta^*\right) \pi_t^*(\theta^* \mid \tilde{\theta}_t^*) \mathrm{d}\theta^* \right\}^2 \tilde{\pi}_t^*(\tilde{\theta}_t^*). \qquad (28)
\end{aligned}
$$

Combining (27) and the lower bound (28) we realise a bound on the information ratio, as below. We have,

$$
\begin{aligned}
\Gamma_t(\tilde{\theta}_t^*, \tilde{\theta}_t) &\leq \frac{\left(\max(l_{01}, 1 - l_{11}) \sum_{\tilde{\theta}_t^*} \left\{ \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^*) \mathrm{d}\theta^* - \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^* \mid \tilde{\theta}_t^*) \mathrm{d}\theta^* \right\} \tilde{\pi}_t^*(\tilde{\theta}_t^*)\right)^2}{2\pi_{t-1}(\Theta_t) \sum_{\tilde{\theta}_t^*} \left\{ \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^*) \mathrm{d}\theta^* - \int \sigma(x_t^\top \theta^*) \pi_t^*(\theta^* \mid \tilde{\theta}_t^*) \mathrm{d}\theta^* \right\}^2 \tilde{\pi}_t^*(\tilde{\theta}_t^*)} \\
&\leq \frac{\max(l_{01}, 1 - l_{11})^2}{2\pi_{t-1}(\Theta_t)},
\end{aligned}
$$

where the final inequality holds by Cauchy-Schwarz. □

D Proof of Proposition 3

Note that for a given θ', and round $t \in [T]$ there are at most three values of the distortion rate. We have, for $\theta' \in \Theta$, and $t \in [T]$,

$$d_t(\theta, \theta') = \begin{cases} 0, & \text{if } \alpha_t(\theta') = \alpha_t(\theta) \\ (1 + l_{01} - l_{11})\sigma(x_t^\top \theta') - 1, & \text{if } \alpha_t(\theta') = 0 \text{ and } \alpha_t(\theta) = 1 \\ 1 - (1 + l_{01} - l_{11})\sigma(x_t^\top \theta'), & \text{if } \alpha_t(\theta') = 1 \text{ and } \alpha_t(\theta) = 0. \end{cases}$$

The condition (9) is equivalent to

$$\max_{t \in [T]} d_t(\theta, \theta') \leq \epsilon, \ \theta, \theta' \in \Theta_k, \ \forall k \in [K].$$

Since d_t depends on the round t only through x_t, we may replace the above condition with the following,

$$d_{\mathcal{X}}(\theta, \theta') := \max_{x \in \mathcal{X}} d(\theta, \theta'; x) \leq \epsilon, \ \theta, \theta' \in \Theta_k, \ \forall k \in [K]. \tag{29}$$

Then any partition satisfying (29) satisfies condition (9) for all T.

To identify the size of such a partition we consider the form of $d_{\mathcal{X}}$. Recall that $d(\theta, \theta'; x) \neq 0$ only when $\alpha(\theta; x) \neq \alpha(\theta'; x)$, thus the context vector $x \in \mathcal{X}$ achieving $\max_{x \in \mathcal{X}} d(\theta, \theta'; x)$ must be such that θ and θ' select different classifications.[7] Thus, we may write

$$d_{\mathcal{X}}(\theta, \theta') := \max_{x \in \mathcal{X}: \alpha(\theta; x) \neq \alpha(\theta'; x)} d(\theta, \theta'; x) = \max_{x \in \mathcal{X}: \alpha(\theta; x) \neq \alpha(\theta'; x)} \left| 1 - (1 + l_{01} - l_{11})\sigma\left(x^\top \theta'\right) \right|.$$

For a parameter vector $\theta \in \Theta$, define the set of contexts on the classification boundary, $\mathcal{X}_\theta \subset \mathcal{X}$, as

$$\mathcal{X}_\theta = \left\{ x \in \mathcal{X} : x^\top \theta = \log\left(\frac{1}{l_{01} - l_{11}}\right) \right\},$$

(note: $\sigma^{-1}\left(1/(1 + l_{01} - l_{11})\right) = \log(1/(l_{01} - l_{11}))$). We then have for $\theta \in \Theta$ fixed that

$$d_{\mathcal{X}}(\theta, \theta') \leq d_{\mathcal{X}_\theta}(\theta, \theta') := \max_{x \in \mathcal{X}_\theta} \left| 1 - (1 + l_{01} - l_{11})\sigma\left(x^\top \theta'\right) \right|.$$

It follows that $d_{\mathcal{X}}(\theta, \theta') \leq \epsilon$ for all θ' such that for all $x \in \mathcal{X}_\theta$

$$x^\top \theta' \in \left[\sigma^{-1}\left(\frac{1 - \epsilon}{1 + l_{01} - l_{11}}\right), \sigma^{-1}\left(\frac{1 + \epsilon}{1 + l_{01} - l_{11}}\right) \right] = \left[\log\left(\frac{1 - \epsilon}{l_{01} - l_{11} + \epsilon}\right), \log\left(\frac{1 + \epsilon}{l_{01} - l_{11} - \epsilon}\right) \right],$$

i.e. over the same range that $d_{\mathcal{X}_\theta}(\theta, \theta') \leq \epsilon$. Expressing this in terms of the x-weighted norm between θ and θ' we can equivalently say that, $d_{\mathcal{X}}(\theta, \theta') \leq \epsilon$ for all θ' such that for all $x \in \mathcal{X}_\theta$,

$$x^\top (\theta - \theta') \in \left[\log\left(\frac{(l_{01} - l_{11}) - \epsilon}{(l_{01} - l_{11}) + (l_{01} - l_{11})\epsilon}\right), \log\left(\frac{(l_{01} - l_{11}) + \epsilon}{(l_{01} - l_{11}) - (l_{01} - l_{11})\epsilon}\right) \right]$$

[7] The only exception is in trivial settings where x is so extreme that there is a single optimal action for all $\theta \in \Theta$.

$$= \left[\log\left(\frac{1 - \frac{\epsilon}{l_{01} - l_{11}}}{1 + \epsilon} \right), \ \log\left(\frac{1 + \frac{\epsilon}{l_{01} - l_{11}}}{1 - \epsilon} \right) \right],$$

and thus also for all θ' such that for all $x \in \mathcal{X}_\theta$,

$$\begin{aligned} \left| x^\top (\theta - \theta') \right| &\leq \min \left\{ \log\left(\frac{1 + \frac{\epsilon}{l_{01} - l_{11}}}{1 - \epsilon} \right), \ -\log\left(\frac{1 - \frac{\epsilon}{l_{01} - l_{11}}}{1 + \epsilon} \right) \right\} \\ &= \log\left(1 + \frac{\epsilon}{\min(1, l_{01} - l_{11})} \right) - \log\left(1 - \frac{\epsilon}{\max(1, l_{01} - l_{11})} \right). \end{aligned} \tag{30}$$

The minimum can be shown to be defined as such by considering the difference

$$\log\left(\frac{1 + \frac{\epsilon}{l_{01} - l_{11}}}{1 - \epsilon} \right) - \left(-\log\left(\frac{1 - \frac{\epsilon}{l_{01} - l_{11}}}{1 + \epsilon} \right) \right) = \log\left(\frac{1 - \frac{\epsilon^2}{(l_{01} - l_{11})^2}}{1 - \epsilon^2} \right),$$

and observing that it is negative when $(l_{01} - l_{11}) < 1$, for $\epsilon > 0$.

We next move from the logarithmic bound (30) to a bound that is linear in ϵ. From the well-known logarithm inequalities,

$$\frac{x}{1 + x} < \log(1 + x) < x, \quad x > 1$$

we also have for $a > 0$ that,

$$\begin{aligned} \frac{ax}{1 + ax} < \log(1 + ax) < ax, \quad x > -1/a, \\ \text{and } \frac{-ax}{1 - ax} < \log(1 - ax) < -ax, \quad x < 1/a. \end{aligned}$$

Defining $l_{min} = \min(l_{01}, 1 - l_{11})$ and $l_{max} = \max(l_{01}, 1 - l_{11})$, it follows that for $\epsilon \in [0, l_{max}^{-1})$,

$$\log\left(1 + \frac{\epsilon}{l_{min}} \right) - \log\left(1 - \frac{\epsilon}{l_{max}} \right) > \frac{\epsilon / l_{min}}{1 + \epsilon / l_{min}} + \frac{\epsilon}{l_{max}} > \frac{\epsilon}{l_{max}}. \tag{31}$$

Thus by the combination of (30) and (31), for a given $\theta \in \Theta$, we have that $d_{\mathcal{X}}(\theta, \theta') \leq \epsilon$ for all $\theta' \in \Theta$ such that

$$\sum_{i=1}^{d} |\theta_i - \theta'_i| \leq \frac{\epsilon}{\max_{x \in \mathcal{X}_\theta} ||x|| \cdot l_{max}}.$$

It follows that the size of a partition satisfying condition (9) may be bounded by the size of an $\epsilon / (l_{max} \max_{x \in \mathcal{X}} ||x||)$-cover of Θ with respect to the ℓ_1 norm. Since $\Theta \subset B_1^d$, the size of the cover of Θ is itself bounded by the size of the cover of B_1^d, and therefore we have,

$$K \leq \left(\frac{3 l_{max} x_{max}}{\epsilon} \right)^d,$$

via the standard result (see e.g. Lemma 1 of Lorentz (1966)) that a δ-cover of a unit ball in d dimensions is of size $(3/\delta)^d$. □

E Polya-Gamma Gibbs Sampler

Following t rounds, where θ^* has prior π, the posterior distribution on θ^* is as follows,

$$\pi(\theta \mid \mathcal{H}_t) = \frac{\pi(\theta)}{D_t} \prod_{s\in[t]:A_s=1} \Big(\sigma(x_s^\top \theta)\Big)^{C_s} \Big(1-\sigma(x_s^\top \theta)\Big)^{1-C_s}, \ \theta \in \Theta, \tag{32}$$

where D_t is the normalising constant,

$$D_t = \int_\Theta \pi(\theta) \prod_{s\in[t]:A_s=1} \Big(\sigma(x_s^\top \theta)\Big)^{C_s} \Big(1-\sigma(x_s^\top \theta)\Big)^{1-C_s} d\theta. \tag{33}$$

Regardless of the choice of prior π, the posterior in (32) is intractable, in the sense that samples cannot readily be drawn directly from it. However, if π is chosen as a multivariate Gaussian then a highly efficient approximate sampling scheme is achievable via augmentation of the likelihood with PG random variables.

A real-valued random variable follows a PG distribution with parameters $b > 0$ and $c \in \mathbb{R}$, $X \sim PG(b,c)$, if

$$X = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{G_k}{(k-\frac{1}{2})^2 + \frac{c^2}{4\pi^2}},$$

where G_k are i.i.d. $Gamma(b,1)$ random variables. Key to the augmentation scheme is the following identity of Polson et al. (2013), for $a \in \mathbb{R}$, $b > 0$ and ω following a $PG(b,0)$ distribution

$$\frac{(e^z)^a}{(1+e^z)^b} = 2^{-b} e^{(a-b/2)z} \int_0^\infty e^{-\omega z^2/2} p(\omega) \mathrm{d}\omega, \quad z \in \mathbb{R}.$$

Applying this identity to the likelihood component of (32), we may write

$$\pi(\theta \mid \mathcal{H}_t) \propto \pi(\theta) \prod_{s\in[t]:A_s=1} \exp\Big((1/2-C_s)x_s^\top \theta\Big) \int_0^\infty \exp\Big(-\omega_s (x_s^\top \theta)^2/2\Big)\, p(\omega_s)\mathrm{d}\omega_s,$$

where each ω_s is a $PG(1,0)$ random variable. It follows that if $\pi(\theta)$ is chosen as a multivariate Gaussian, s.t. $\theta^* \sim MVN(\mathbf{b}, \mathbf{B})$, the posterior on θ^*, conditioned on PG random variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{|\ldots|})$, is also multivariate Gaussian,

$$\pi(\theta \mid \boldsymbol{\omega}, \mathcal{H}_t) \propto \pi(\theta) \prod_{s\in[t]:A_s=1} \exp\left(\frac{\omega_s}{2}\left(x_s^\top \theta - \frac{(1-C_s)}{\omega_s}\right)^2\right).$$

As such, we may construct a Gibbs sampler, which iterates between sampling PG random variables, and from the conditional Gaussian on θ^*. Sampling of PG random variables is highly efficient, due to a rejection sampler of Polson et al. (2013) with acceptance probability no less than 0.9992.

Our PG-TS approach (including the Gibbs steps) is summarised in Algorithm 1. It uses an additional counter random variable $N(t) = \sum_{s=1}^t \mathbb{I}\{A_s = 1\}$ to track the number of rounds in which the class 1 has been chosen, and assumes (in line with the definition of the model) that C_t can be recovered from $\Phi_t(1)$. Further, for $n \in \mathbb{N}$ it uses $R(n) = \min(t \geq n : N(t) = n)$ to refer to the round in which class 1 is chosen for the n^{th} time, and $\mathbf{X}_n$ to represent the matrix whose columns are the feature vectors $x_{R(1)}, x_{R(2)}, \ldots, x_{R(n)}$ (in that order).

In each round PG-TS draws M via the Gibbs sampling possible due to PG augmentation. Here we describe the simplest version of the algorithm, in the sense that a fixed number of samples, M, are used to estimate the posterior in each round, but a time dependent $M(t)$ could also be used.

Inputs: Prior mean vector $\mathbf{b}$, Prior precision matrix $\mathbf{B}$, Number of Gibbs iterations M, Observed classes $\mathbf{C} = (C_1, \ldots, C_n)$, Context matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, Initialisation parameter $\theta^{(0)}$

Compute $\boldsymbol{\kappa} = (C_1 - \frac{1}{2}, \ldots, C_n - \frac{1}{2})$
for $m = 1$ **to** M **do**
  for $i = 1$ **to** n **do**
    Draw $\omega_i \mid \theta^{(m-1)} \sim PG(1, x_i^\top \theta^{(m-1)})$.
  end for
  Compute $\Omega = \text{diag}(\omega_1, \ldots, \omega_n)$
  Compute covariance matrix $V_{\boldsymbol{\omega}} = (\mathbf{X}^\top \Omega \mathbf{X} + \mathbf{B}^{-1})^{-1}$
  Compute mean vector $m_{\boldsymbol{\omega}} = V_{\boldsymbol{\omega}}(\mathbf{X}^\top \boldsymbol{\kappa} + \mathbf{B}^{-1}\mathbf{b})$
  Draw $\theta^{(m)} \mid \boldsymbol{\kappa}, \boldsymbol{\omega} \sim MVN(m_{\boldsymbol{\omega}}, V_{\boldsymbol{\omega}})$
end for
return $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$

Algorithm 3: GIBBS

F Proof of Theorem 2

The proof utilises uniform ergodicity of the Gibbs sampler, together with convergence of the posterior to demonstrate that the α-divergence is vanishing in the limit.

Choi and Hobert (2013) have shown that the PG-Gibbs sampler is uniformly ergodic, and thus as the number of Gibbs samples M approaches ∞, the distribution $\pi_t^{(M)}$ shows convergence to π_t. Specifically, that there exists $\rho \in [0, 1)$ such that,

$$\left|\pi_T - \pi_T^{(M)}\right|_{TV} \leq \left|\pi_T - \pi_T^{(0)}\right|_{TV} \rho^M,$$

where $|\cdot|_{TV}$ denotes the total variation distance between probability distributions. Here, however, we consider an analysis of the finite M, infinite T setting, so this result alone does not imply shrinkage of the α-divergence.

To prove such a result, we must show that TS using either of π_t or $\pi_t^{(M)}$ will lead to using the informative action (i.e. $A_t = 1$) infinitely often. In either case, the draws of samples θ_t from π_{t-1} or $\theta_t^{(M)}$ from $\pi_{t-1}^{(M)}$ are independent of the draw of a context x_t from p_X. As such, by Assumption 1 we have $\mathbb{E}_{t-1}(p_X(\mathcal{X}_1(\theta_t))) > \delta$ and $\mathbb{E}_{t-1}^{(M)}(p_X(\mathcal{X}_1(\theta_t^{(M)}))) > \delta$, for all t. This provides the following guarantees on the rate of selection of the informative action,

$$\lim_{T\to\infty} \mathbb{E}_0 \left(\frac{\sum_{t=1}^T \mathbb{I}\{A_t^{TS} = 1\}}{T} \right) > \delta, \text{ and} \tag{34}$$

$$\lim_{T\to\infty} \mathbb{E}_0 \left(\frac{\sum_{t=1}^T \mathbb{I}\{A_t^{(M)} = 1\}}{T} \right) > \delta. \tag{35}$$

It therefore follows that under either sampling regime (i.e. either the exact or approximate TS algorithms) the number of observed labels approaches infinity as T does. As such, the posterior

induced under either regime is strongly consistent (Ghosal et al., 1995, Proposition 1) and satisfies the following stationary convergence guarantee (Yang and Dunson, 2013, Lemma 3.7) with respect to the L_1 norm:

$$|\pi_T - \pi_{T-1}|_1 \to 0, \text{ as } T \to \infty.$$

Now fix $\epsilon > 0$. Since the TV norm is bounded by half the L_1 norm between densities (when the densities exist — see e.g. equation (1) of Devroye et al. (2018)), there exists $S > 0$ such that $|\pi_t - \pi_{t-1}|_{TV} < \epsilon(1-\rho^M)/(3\rho^M)$ for all $t > S$. Hence, we have, by repeated application of the triangle inequality and the uniform ergodicity result, that,

$$\begin{aligned}
|\pi_T - \pi_T^{(M)}|_{TV} &\leq |\pi_T - \pi_T^{(0)}|_{TV}\rho^M \\
&\leq |\pi_T - \pi_{T-1}|_{TV}\rho^M + |\pi_{T-1} - \pi_T^{(0)}|_{TV}\rho^M \\
&= |\pi_T - \pi_{T-1}|_{TV}\rho^M + |\pi_{T-1} - \pi_{T-1}^{(M)}|_{TV}\rho^M \\
&\leq \sum_{t=1}^{T} |\pi_{T+1-t} - \pi_{T-t}|_{TV}\rho^{tM} + |\pi_0 - \pi_0^{(M)}|_{TV}\rho^{MT} \\
&= \sum_{t=1}^{T-S} |\pi_{T+1-t} - \pi_{T-t}|_{TV}\rho^{tM} \\
&\qquad + \sum_{t=T-S+1}^{T} |\pi_{T+1-t} - \pi_{T-t}|_{TV}\rho^{tM} + |\pi_0 - \pi_0^{(M)}|_{TV}\rho^{MT} \\
&< \frac{\epsilon(1-\rho^M)}{3\rho^M} \sum_{t=1}^{T-S} \rho^{tM} + \sum_{t=T-S+1}^{T} \rho^{tM} + \rho^{MT} \\
&< \frac{\epsilon(1-\rho^M)}{3\rho^M} \sum_{t=1}^{\infty} \rho^{tM} + \rho^{(T-S)M} \sum_{t=1}^{\infty} \rho^{tM} + \rho^{TM} \\
&= \frac{\epsilon}{3} + \rho^{(T-S)M} \frac{\rho^M}{1-\rho^M} + \rho^{TM}
\end{aligned}$$

For T sufficiently large that $\rho^{(T-S)M} < (1-\rho^M)\epsilon/(3\rho^M)$ and $\rho^{TM} < \epsilon/3$, we see that $|\pi_T - \pi_T^{(M)}|_{TV} < \epsilon$. Since $\epsilon > 0$ is arbitrary, we see that $|\pi_T - \pi_T^{(M)}|_{TV} \to 0$ as $T \to \infty$. □

G Proof of Theorem 3

The proof in this section is somewhat informal, but constructed as such with the intention of limiting simple but lengthy translations of existing results to near-identical settings. Ultimately, the proof amounts to demonstrating that since PG-TS will select the informative action infinitely often, and since the posterior approximation will converge to the underlying true parameter θ^*, the proportion of rounds in which PG-TS selects the action which is optimal in expectation approaches 1 as the number of rounds approaches infinity.

A similar result is established in an alternative, and more traditional, contextual bandit setting by May et al. (2012), in their Theorem 1. Therein each action $a \in \mathcal{A}$ (where $|\mathcal{A}|$ may be greater than 2) is associated with a continuous reward function $f_a : \mathcal{X} \to \mathbb{R}$, and each of these functions may have separate parameters. May et al. (2012) demonstrate that an asymptotic consistency

result is enjoyed by any TS-like algorithm for such a problem subject to conditions on the sampling distribution. Therein a TS-like algorithm is defined as one which samples reward functions $\tilde{f}_{t,a}$ from distributions $Q_{t,a}$ at each time t, and plays the action with the largest $\tilde{f}_{t,a}(x_t)$ value, and sufficient conditions are expressed in terms of the distributions $Q_{t,a}$.

The basis of our informal proof in this section is to demonstrate that LCAT is sufficiently similar to the aforementioned contextual bandit problem, and the approximate posterior distributions used by PG-TS satisfy appropriate conditions such that the asymptotic consistency result can be extended to this setting.

Two conditions are critical to the asymptotic consistency result in May et al. (2012). First, where $n_{t,a} = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$ is defined to be the number of plays of action a in t rounds, that we have

$$\mathbb{P}\left(\cup_{a \in \mathcal{A}} \left\{n_{t,a} \to \infty \text{ as } t \to \infty\right\}\right), \tag{36}$$

i.e. that every action is sampled infinitely often in the limit. Second, that for each action $a \in \mathcal{A}$, we have convergence of the sampling distribution to the true reward function, i.e.

$$\left[Q_{t,a} - f_a(x_t)\right] \to^{\mathbb{P}} 0 \text{ as } n_{t,a} \to \infty. \tag{37}$$

These conditions are the ultimate (and critical) consequence of Assumptions 1, 2, and 4, and the intermediate Lemma 2 in May et al. (2012).

Proceeding, we first show that the LCAT problem may equivalently be viewed as a contextual bandit problem with $\mathcal{A} = 2$, and a known reward function for one action. Although this was a step we avoided in the main Bayesian regret analysis, since a bespoke analysis was more powerful than relying on generic contextual bandit results, it is nevertheless useful in this instance, to obtain the consistency result for the approximate algorithm.

We recall the form of the loss and signal matrices used previously,

$$\mathbf{L} = \begin{pmatrix} 0 & l_{01} \\ 1 & l_{11} \end{pmatrix} \text{ and } \mathbf{\Phi} = \begin{pmatrix} 0 & 0 \\ 1 & l_{11} \end{pmatrix},$$

and define shifted versions of these,

$$\mathbf{L}' = \mathbf{L} - \begin{pmatrix} 0 & l_{01} \\ 0 & l_{01} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & l_{11} - l_{01} \end{pmatrix} \text{ and } \mathbf{\Phi}' = \begin{pmatrix} 0 & 0 \\ 1 & l_{11} - l_{01} \end{pmatrix}.$$

The shifted loss matrix has merely changed the scale of the losses for the event $C_t = 1$, and the shifted signal matrix replaces one (essentially) arbitrary signal l_{11} with another $l_{11} - l_{01}$.

The contextual partial monitoring problem characterised by loss matrix $\mathbf{L}'$ and signal matrix $\mathbf{\Phi}'$ is recognised as a 2-armed logistic contextual bandit. In particular, taking rewards to be the negative of losses, the action indexed 0 has expected reward function

$$f_0(x) = 0, \quad x \in \mathcal{X}$$

and the action indexed 1 has expected reward function

$$f_1(x) = (1 + l_{01} - l_{11})\sigma(x^\top \theta) - 1, \quad x \in \mathcal{X}.$$

The reward observations have zero noise under selection of action 0, and are supported on $\{-1, l_{01} - l_{11}\}$ under selection of action 1. A straightforward adaptation of the action selection step implements a version of PG-TS for this version of the problem using the same Gibbs sampler and structure.

Having framed the LCAT problem as a contextual bandit problem, proof of the asymptotic consistency result then reduces to the verification of conditions (36) and (37). Firstly, we have from (35) that PG-TS will sample the informative action infinitely often, and by an equivalent proof under Assumption 1 the same is true of the non-informative action. Thus we have that (36) is satisfied by PG-TS. Plainly, (37) holds for the non-informative action since its rescaled reward function is known by design, and (37) is shown to hold for the informative action by the consistent estimation result of Theorem 2. Thus, by extension of the results of May et al. (2012), we have the asymptotic consistency and asymptotically sublinear regret of the PG-TS algorithm. □

H Pseudocode for CBP-SIDE Algorithm for Apple Tasting

In this section we provide the particular version of the more general CBP-SIDE algorithm used for the contextual logistic apple tasting problem. Due to the small action set, and specific loss model, the statement of the algorithm can be streamlined. That said, the correct choice of estimator, confidence set, and various constants is still non-trivial. We explain what we believe to be the best theoretically-supported choice of these components in this section.

CBP-SIDE requires the identification of *observer vectors*, v_{ij} and v_{ji}, for each pair of actions, i, j (in our case, the only action pairs are of course $\{0,0\}$, $\{0,1\}$, and $\{1,1\}$), which satisfy,

$$l_i - l_j = S_i^\top v_{ij} - S_j^\top v_{ji},$$

where l_i and l_j are columns of the loss matrix L and S_i and S_j are *signal matrices* - which are exactly the incidence matrices of the symbols in columns of Φ. In our setting, we have $l_0 = (0\ 1)$, $l_1 = (l_{01}\ l_{11})$ and signal matrices,

$$S_0 = \begin{pmatrix} 1 & 1 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

It is clear that valid v_{ij} and v_{ji} are not unique, but Bartók and Szepesvári (2012) note that a theoretically optimal choice, is

$$\begin{pmatrix} v_{ij} \\ -v_{ji} \end{pmatrix} = \begin{pmatrix} S_i^\top & S_j^\top \end{pmatrix}^+ (l_i - l_j).$$

So in our case we may choose scalar $v_{01} = -l_{01}$ and $v_{10} = l_{11} - 1$, and $v_{00} = v_{11} = 0$.

In each round, the general CBP-SIDE algorithm computes a estimate $\hat{\theta}$ of the unknown parameter, and from this a regret term Δ_{ij} and a confidence term c_{ij} for each pair of actions i, j, of the form

$$\begin{aligned} \Delta_{ij} &= v_{ij}\hat{q}_i + v_{ji}\hat{q}_j \\ c_{ij} &= |v_{ij}|w(i,t) + |v_{ji}|w(j,t), \end{aligned}$$

where $\hat{q}_k$ is the estimated probability action k is suboptimal and $w(\cdot,\cdot)$ is a confidence width function chosen to establish theoretical guarantees. Since we choose $v_{00} = v_{11}$ we can focus solely on the case $\{i,j\} = \{0,1\}$ and compute a single regret term

$$\hat{\Delta} := \Delta_{01} = -l_{01}\left(1 - \sigma\left(x_t^\top \hat{\theta}\right)\right) + (l_{11} - 1)\sigma\left(x_t^\top \hat{\theta}\right),$$

and confidence term

$$\hat{c} := c_{01} = (1 + l_{01} - l_{11})\, w(t),$$

with w specified below.

We use the maximum likelihood estimator, projected back in to Θ as our estimator of θ^*. Specifically, with respect to the matrix

$$V_t = \sum_{s=1}^{t-1} \mathbb{I}\{A_t = 1\} x_s^\top x_s,$$

and the log-likelihood maximised over $\mathbb{R}^d$ at θ_t^{MLE}, we define the estimator

$$\hat{\theta}_t = \underset{\theta \in \Theta}{\operatorname{argmin}} \, ||\theta - \theta^{MLE}||^2_{V_t^{-1}}.$$

We adapt the confidence width function given in Bartók and Szepesvári (2012) for online multinomial logistic regression, leading to the following expression derived from the self-normalised inequalities of Abbasi-Yadkori et al. (2011),

$$w(t) = C \left(\sqrt{2d(1 + N(t-1)R^2/d) + 2\log(1/\delta_{N(t-1)})} + Rd \right) \sqrt{x_t^\top V_t^{-1} x_t}.$$

Here, the $\Theta-$ and $\mathcal{X}-$dependent constant $C = [\inf_{\theta \in \Theta, x \in \mathcal{X}} (1 - \sigma(x^\top \theta))\sigma(x^\top \theta)]^{-1}$ arises from Lemma 3 of Bartók and Szepesvári (2012), R is a bound on the 2-norm of the features $x \in \mathcal{X}$, $N(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = 1\}$ counts the number of uses of action 1 in t rounds (i.e. the size of the observed data), and $\delta_s = s^{-2}$ is chosen to realise an optimal regret bound.

Finally, the action selection process also simplifies with respect to the general case, and we assign class 1, unless the regret term $\hat{\Delta}$ falls below $-\hat{c}$, indicating that a prediction of class 0 with sufficient confidence to avoid pass on the information gaining action. Algorithm 4 gives the adaptation of CBP-SIDE to apple tasting, incorporating the above specifications.

Inputs: Loss parameters l_{01}, l_{11}
Initialise: $\mathcal{D} = \emptyset$.
for $t = 1, 2, \ldots$ **do**
  Receive context $x_t \in \mathbb{R}^d$
  Compute regret estimate $\hat{\Delta} = (1 + l_{01} - l_{11})\sigma(x_t \hat{\theta}_t) - l_{01}$
  Compute confidence width $\hat{c} = (1 + l_{01} - l_{11}) w(t)$
  Select action $A_t = 1 - \mathbb{I}\{\hat{\Delta} \leq -\hat{c}\}$
  if $A_t = 1$ **do**
    Observe $\Phi_t(1) \in \{1, l_{11}\}$
    Augment $\mathcal{D} \leftarrow \mathcal{D} \cup \{x_t, \Phi_t(1)\}$
  end if
end for

Algorithm 4: CBP-SIDE for Apple Tasting

I Parameter Tuning for PG-IDS

In this section we give the results of further experiments used to estimate the optimal parameter λ for the PG-IDS scheme. We consider ten choices of $\lambda = \{0, 0.05, \ldots, 0.45\}$ and compare the performance of the resulting PG-IDS algorithms with each other and PG-TS (with the same prior). To identify a robust choice of parameter we investigate both problem (a) and problem (b) from the main text, but with a range of horizons, loss matrices and context distributions.

For completeness, recall that problem (a) has $\theta^* = (0.5, 0.9, -0.75)$ and contexts sampled from a mixture of multivariate Gaussian with independent components whose variances are 0.2. With probability p_{mix} the mean vector is $(0, 0, 0.875)$ and with probability $1 - p_{mix}$ it is $(0.875, 0.2, 0)$. We construct different variants of problem (a) by varying l_{01}, and p_{mix}, as well as the problem horizon, T. We will keep $l_{11} = 0.05$ fixed.

We run PG-IDS for the ten λ values and PG-TS on nine variants of problem (a) per time horizon value. We index the mixing parameters 1. $p_{mix} = 0.95$, 2. $p_{mix} = 0.7$, 3. $p_{mix} = 0.5$, and the loss values, i. $l_{01} = 0.5$, ii. $l_{01} = 0.7$, iii. $l_{01} = 0.95$. So problem (a).1.i. for instance, refers to the variant of problem (a) with $p_{mix} = 0.95$, and $l_{01} = 0.5$. This indexing is to make graphical presentation of the results more straightforward.

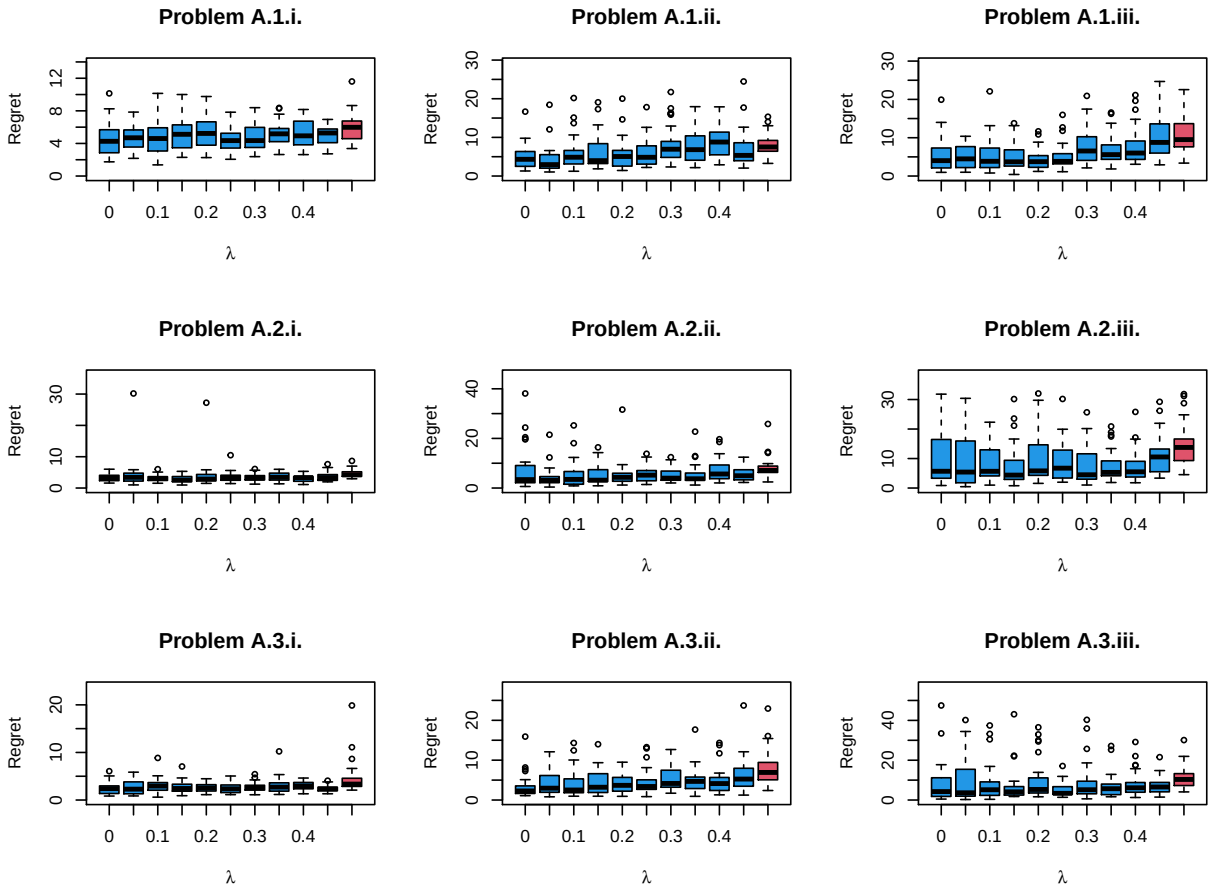


Figure 3: Plots showing the effects of problem and algorithm parameters on regret, with $T = 250$. In each plot, blue boxplots show the distribution of cumulative regret for tunable PG-IDS with tuning parameter in $\{0, 0.05, 0.1, \ldots, 0.45\}$ and the red (rightmost) boxplots show the distribution of cumulative regret for PG-TS for comparison.

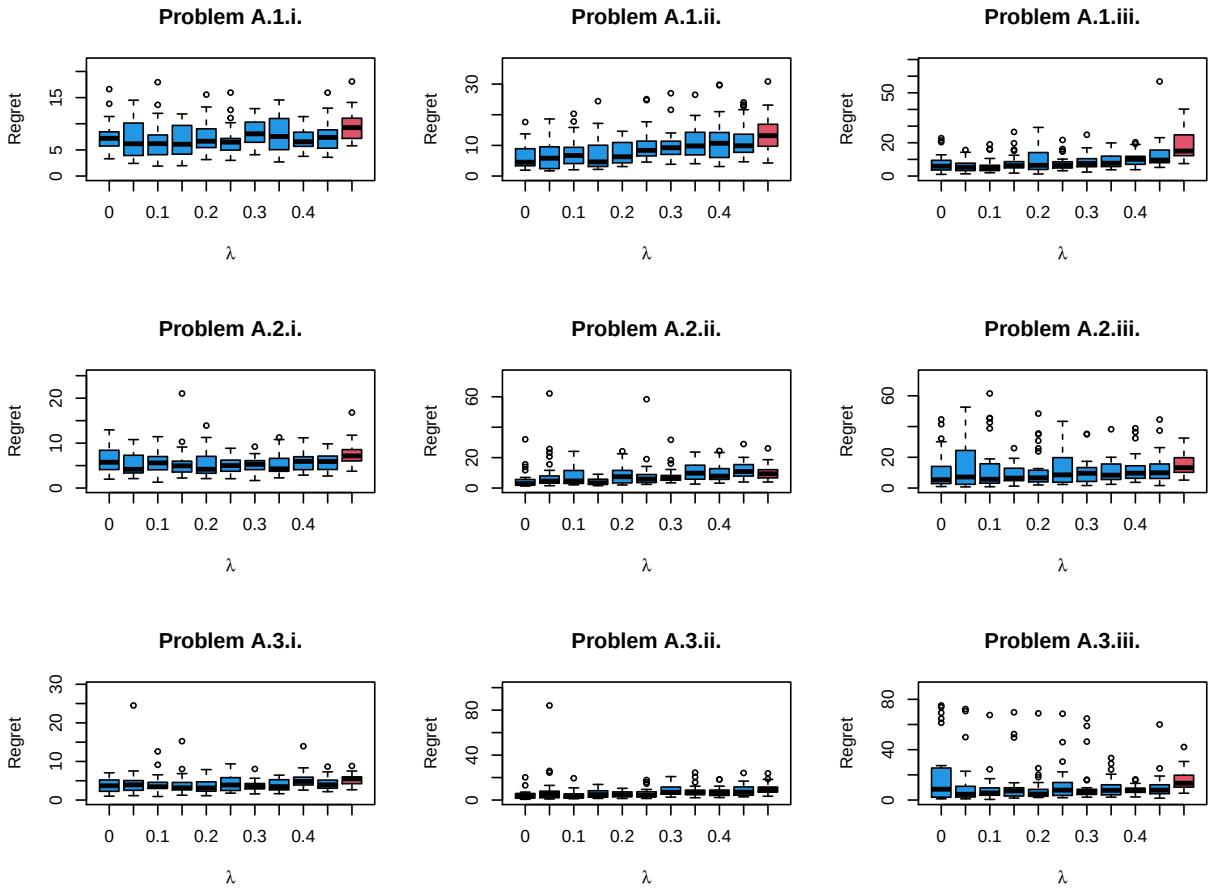


Figure 4: Plots showing the effects of problem and algorithm parameters on regret, with $T = 500$. In each plot, blue boxplots show the distribution of cumulative regret for tunable PG-IDS with tuning parameter in $\{0, 0.05, 0.1, \ldots, 0.45\}$ and the red (rightmost) boxplots show the distribution of cumulative regret for PG-TS for comparison.