Thompson Sampling for Noisy Contextual Bandits with Delayed Observations: An Information-Theoretic Regret Analysis

Sharu Theresa Jose and Shana Moothedath

***Abstract*— We study stochastic linear contextual bandits (CB) where the agent observes a *noisy* version of the true context through a noise channel with unknown channel parameter. The agent chooses an action based on the observed noisy context and receives the corresponding reward. The exact context is disclosed to the agent after a delay, subsequent to the action selection. Our objective is to design an action policy that can "approximate" that of a Bayesian oracle that has access to the reward model and the noise channel parameter. We introduce a modified, Bayesian Thompson sampling algorithm and analyze its Bayesian cumulative regret with respect to the oracle action policy via information-theoretic analysis. For d-dimensional Gaussian bandits with Gaussian context noise, our information-theoretic analysis shows that the Bayesian cumulative regret scales as $\tilde{O}(d\sqrt{T})$, where T is the horizon. We empirically evaluate the performance of our algorithms against baselines.**

***Index Terms*— Multi-armed bandits, online learning, Thompson Sampling, Information-theoretic analysis**

I. Introduction

Multi-Armed Bandit (MAB) deals with uncertain decision-making problems, where an agent acquires the ability to make optimal decisions through repetitive interactions with a stochastic environment. They find applications across various domains such as control and robotics [1], [2], clinical trials [3], communications [4], and smart grids [5]. A widely popular MAB framework is contextual bandits (CBs), which captures the essence of sequential decision-making by incorporating side information, termed *context* [6]. In the standard CB model, an agent interacts with the environment over numerous rounds. In each round, the environment presents a context to the agent based on which the agent chooses an action and receives a reward from the environment. The goal of the agent is to design a policy for action selection that can maximize the cumulative mean reward accrued over a T-length horizon.

While most prior research on CBs has primarily focused on models with known exact contexts [7]–[9], in many real-world applications, the observed contexts are noisy due to imprecise measurements or predictions. For instance, in robotics applications, the context, such as sensory data from cameras/sensors, may be subject to noise due to unpredictable factors like varying lighting, environmental obstructions or sensor inaccuracies. In such scenarios, when the exact contexts are unknown, the agent must utilize the observed noisy contexts to estimate the mean reward associated with the true context. However, this results in a biased estimate that renders the application of standard CB algorithms unsuitable.

S. T. Jose is with Department of Computer Science, University of Birmingham, Birmingham -B15 2TT, UK s.t.jose@bham.ac.uk

S. Moothedath is with the Department of Electrical Engineering, Iowa State University, Ames, IA 50011, USA mshana@iastate.edu

We consider a CB problem with noisy context observations. In each round, the environment samples a true context vector c_t from a *context distribution* that is *known* to the agent. The agent only observes a noisy context $\hat{c}_t$ obtained as the output of a noise channel $P(\hat{c}_t|c_t, \gamma^*)$ parameterized by γ^*. The agent does not know the channel parameter γ^* and the true context c_t. Based on the observed noisy contexts, the agent chooses an action a_t and observes a reward r_t corresponding to the true context. After the reward is revealed, the agent observes the true context c_t with a delay. Such scenarios occur when decision-making is influenced by noisy signals or data, and the accurate measurements are only revealed with some delay afterward, as in remote sensing systems.

We consider a linear CB whose mean reward $\phi(a_t, c_t)^\top \theta^*$ is determined by an unknown reward parameter θ^* and the feature vector $\phi(a_t, c_t)$. The goal of the agent is to design an action policy that minimizes the *Bayesian cumulative regret* with respect to the action policy of a Bayesian oracle. The oracle has access to the reward model and the channel parameter γ^*, and uses the predictive distribution of the true context given the observed noisy context to select an action.

Related Works: Reference [10] considers a setting where there is a bounded zero-mean noise in the *feature vector* (denoted by $\phi(a, c)$, where a is the action and c is the context) rather than in the context vector, and the agent observes only noisy features. For this setting, they develop an upper confidence bound (UCB) algorithm with frequentist regret that scales as $\tilde{O}(d\sqrt{T})$ where d is the dimension of the feature vector. Reference [11] models the uncertainty regarding the true contexts by a *context distribution* that is known to the agent, while the agent never observes the true context; and develops a UCB algorithm with a regret of the order $\tilde{O}(d\sqrt{T})$. A similar setting has also been considered in [12]. Different from these works, [13] considers the setting where the true feature vectors are sampled from an unknown feature distribution at each time, but the agent observes only a noisy feature vector. Assuming Gaussian feature noise with unknown mean and covariance, they develop an OFUL algorithm with frequentist regret of the order $\tilde{O}(d\sqrt{T})$.

Our Contributions: In this letter, we propose a fully Bayesian TS algorithm that approximates the Bayesian oracle policy. A longer version of this letter where we also study a fully unobserved true context setting can be found in the arXiv version [14]. Our proposed algorithm differs from the standard contextual TS [9] in the following aspect. Since the true context vectors are not observed at the time of action selection and the channel parameter γ^* is unknown, the agent uses its knowledge of the context distribution

Preprint Received March 8, 2024 16:00:45 Pacific Time

and the past observed noisy and true contexts to infer a *predictive posterior* distribution of the true context from the current observed noisy context. The inferred predictive distribution is then used to choose the action. This *de-noising* step enables our algorithm to 'approximate' the oracle action policy that uses knowledge of the channel parameter γ^* to implement *exact de-noising*. Moreover, different from existing works that focus on frequentist regret analysis, we derive a novel *information-theoretic* bound on the *Bayesian cumulative regret* of our algorithm. Our main result shows that for Gaussian bandits with Gaussian context noise, the information-theoretic Bayesian regret bound scales as $\tilde{O}(d\sqrt{T})$.

II. Problem Setting

In this section, we present the stochastic linear CB problem studied in this paper. Let $\mathcal{A}$ denote the action set with K actions and $\mathcal{C}$ denote the (possibly infinite) set of d-dimensional context vectors. The environment randomly draws a context vector $c_t \in \mathcal{C}$, at round $t \in \mathbb{N}$, according to a *context distribution* $P(c)$ defined over the space $\mathcal{C}$ of context vectors. $P(c)$ is known to the agent. However, the agent does not observe the true context c_t and only observes a noisy version $\hat{c}_t$, obtained as the output of a noisy, stochastic channel $P(\hat{c}_t|c_t,\gamma^*)$ with the true context c_t as the input, where γ^* denotes the noise channel parameter that is *unknown* to the agent. The agent then chooses an action $a_t \in \mathcal{A}$ according to a (stochastic) action policy $\pi_t(\cdot|\hat{c}_t)$ and receives reward,

$$r_t = \phi(a_t, c_t)^\top \theta^* + \xi_t, \tag{1}$$

where $\phi : \mathcal{A} \times \mathcal{C} \to \mathbb{R}^m$ is the feature map and ξ_t is a zero-mean reward noise. After the agent receives the reward r_t, the environment reveals the true context c_t to the agent. Since the agent observes the true context after a delay, we refer to our problem as *CBs with delayed context observation*. At the end of iteration t, the agent collects the history $\mathcal{H}_{t,r,a,c,\hat{c}} = \{r_\tau, a_\tau, c_\tau, \hat{c}_\tau\}_{\tau=1}^t$ of observed reward-action-context-noisy context tuples. The action policy $\pi_{t+1}(\cdot|\hat{c}_{t+1})$ at $(t+1)$th iteration may depend on the history $\mathcal{H}_{t,r,a,c,\hat{c}}$. The goal of the agent is to devise an action policy that minimizes the *Bayesian cumulative regret* with respect to a baseline action policy. We define Bayesian cumulative regret next.

A. Bayesian Cumulative Regret

The cumulative regret of an action policy $\pi_t(\cdot|\hat{c}_t)$ quantifies how far the mean reward accumulated over T iterations is from that accrued by a baseline action policy $\pi_t^*(\cdot|\hat{c}_t)$. In this work, we consider as baseline the action policy of an *oracle* that has access to the channel noise parameter γ^*, reward parameter θ^*, the context distribution $P(c)$ and the noise channel likelihood $P(c_t|\hat{c}_t,\gamma^*)$. Accordingly, at each iteration t, the oracle can infer the *exact predictive distribution* $P(c_t|\hat{c}_t,\gamma^*)$ of the true context from the observed noisy context $\hat{c}_t$ via Baye's rule as

$$P(c_t|\hat{c}_t,\gamma^*) = \frac{P(c_t,\hat{c}_t|\gamma^*)}{P(\hat{c}_t|\gamma^*)}. \tag{2}$$

Here, $P(c_t,\hat{c}_t|\gamma^*) = P(c_t)P(\hat{c}_t|c_t,\gamma^*)$ is the joint distribution of the true and noisy contexts given the noise channel parameter γ^*, and $P(\hat{c}_t|\gamma^*)$ is the distribution obtained by marginalizing $P(c_t,\hat{c}_t|\gamma^*)$ over the true contexts, *i.e.*,

$$P(\hat{c}_t|\gamma^*) = \mathbb{E}_{P(c_t)}[P(\hat{c}_t|c_t,\gamma^*)], \tag{3}$$

where $\mathbb{E}_\bullet[\cdot]$ denotes expectation with respect to '$\bullet$'. The oracle action policy then adopts an action

$$\begin{aligned} a_t^* &= \arg\max_{a\in\mathcal{A}} \mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a,c_t)^\top\theta^*] \\ &= \arg\max_{a\in\mathcal{A}} \psi(a,\hat{c}_t|\gamma^*)^\top\theta^*, \end{aligned} \tag{4}$$

at iteration t, where $\psi(a,\hat{c}|\gamma^*) := \mathbb{E}_{P(c|\hat{c},\gamma^*)}[\phi(a,c)]$ is the expected feature map with respect to the exact predictive distribution. Note that as in [13], [15], we do not choose the stronger oracle action policy of $\arg\max_{a\in\mathcal{A}} \phi(a,c_t)^\top\theta^*$, that requires access to the true context c_t, as it is generally not achievable by an agent that observes only noisy context $\hat{c}_t$ and has no access to parameter γ^*.

For fixed parameters θ^* and γ^*, we define the cumulative regret of the action policy $\pi_t(\cdot|\hat{c}_t)$ as

$$\mathcal{R}^T(\pi|\theta^*,\gamma^*) = \mathbb{E}\Big[\sum_{t=1}^T \phi(a_t^*,c_t)^\top\theta^* - \phi(a_t,c_t)^\top\theta^* \,\Big|\, \theta^*,\gamma^*\Big], \tag{5}$$

the expected difference in mean rewards of the oracle decision policy and the agent's decision policy over T iterations. The expectation is taken over the randomness in the selection of actions a_t^* and a_t, as well as true context c_t. Importantly, the cumulative regret of (5) can be equivalently written as

$$\begin{aligned} \mathcal{R}^T(\pi|\theta^*,\gamma^*) &= \sum_{t=1}^T \mathbb{E}\big[\mathbb{E}\big[\phi(a_t^*,c_t)^\top\theta^* - \phi(a_t,c_t)^\top\theta^* \big|\hat{c}_t,a_t\big]\big|\theta^*,\gamma^*\big] \\ &= \sum_{t=1}^T \mathbb{E}\big[\psi(a_t^*,\hat{c}_t|\gamma^*)^\top\theta^* - \psi(a_t,\hat{c}_t|\gamma^*)^\top\theta^* \big|\theta^*,\gamma^*\big]. \end{aligned} \tag{6}$$

Our focus in this work is on a *Bayesian framework* where we assume that the reward parameter $\theta^* \in \Theta$ and channel noise parameter $\gamma^* \in \Gamma$ are independently sampled by the environment from prior distributions $P(\theta^*)$, defined on the set Θ of reward parameters, and $P(\gamma^*)$, defined on the set Γ of channel noise parameters, respectively. The agent has knowledge of the prior distributions, the reward likelihood in (1) and the noise channel likelihood $P(\hat{c}_t|c_t,\gamma^*)$, although it does not know the sampled γ^* and θ^*. Using the above prior distributions, we define *Bayesian cumulative regret* of the action policy $\pi_t(\cdot|\hat{c}_t)$ as

$$\mathcal{R}^T(\pi) = \mathbb{E}[\mathcal{R}^T(\pi|\theta^*,\gamma^*)], \tag{7}$$

where the expectation is taken with respect to the prior distributions $P(\theta^*)$ and $P(\gamma^*)$.

In the rest of this letter, we consider multi-variate Gaussian context distribution $P(c) = \mathcal{N}(\mu_c,\Sigma_c)$ with mean $\mu_c \in \mathbb{R}^d$ and covariance $\Sigma_c \in \mathbb{R}^{d\times d}$. The context noise channel $P(\hat{c}|c,\gamma^*)$ is also Gaussian with mean (γ^*+c) and covariance $\Sigma_n \in \mathbb{R}^{d\times d}$. We consider a Gaussian prior distribution $P(\gamma^*) = \mathcal{N}(\mathbf{0},\Sigma_\gamma)$ on the noise channel parameter γ^* with d-dimensional zero mean vector $\mathbf{0}$ and covariance $\Sigma_\gamma \in \mathbb{R}^{d\times d}$.

We assume that Σ_c, Σ_γ and Σ_n are all positive definite matrices. We have the following assumption on our model.

Assumption 2.1: The reward noise ξ_t follows a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with mean 0 and variance $\sigma^2 > 0$. We assume a Gaussian prior $P(\theta^*) = \mathcal{N}(\mathbf{0}, \lambda\mathbb{I})$, with $\mathbb{I}$ denoting identity matrix, on the reward parameter with mean zero and an $m \times m$ diagonal, covariance matrix with entries $\lambda > 0$.

The choice of the Gaussian framework above is due to the easy tractability of posterior and predictive posterior distributions involved in the TS algorithm. We note that similar Gaussian contextual bandit problem with Gaussian context noise has been studied in [13] where they developed an UCB-algorithm that achieves sub-linear frequentist regret.

III. TS Algorithm for Linear-Gaussian Bandits with Delayed True Contexts

A. Preliminaries

In this section, we discuss some key information-theoretic tools that are used to upper bound the Bayesian cumulative regret of (7). To start, let $P(x)$ and $Q(x)$ denote two probability distributions defined over the space $\mathcal{X}$ of random variables x. Then, the Kullback Leibler (KL)-divergence between the distributions $P(x)$ and $Q(x)$ is defined as

$$D_{\mathrm{KL}}(P(x)||Q(x)) = \mathbb{E}_{P(x)}\left[\log \frac{P(x)}{Q(x)}\right], \tag{8}$$

if $P(x)$ is absolutely continuous with respect to $Q(x)$, and it takes value ∞ otherwise. If x and y denote two random variables described by the joint probability distribution $P(x, y)$, then the mutual information (MI) $I(x; y)$ between x and y is defined as $I(x; y) = D_{\mathrm{KL}}(P(x, y)\|P(x)P(y))$, where $P(x)$ (and $P(y)$) is the marginal distribution of x (and y). More generally, for three random variables x, y and z with joint distribution $P(x, y, z)$, the conditional mutual information $I(x; y|z)$ between x and y given z can be expressed as $I(x; y|z) = \mathbb{E}_{P(z)}[D_{KL}(P(x, y|z)\|P(x|z)P(y|z))]$ where $P(x|z)$, $P(y|z)$ are the conditional distributions. We will also use the variational representation of the KL-divergence in the form of the following *Donskar-Varadhan (DV)* inequality,

$$D_{\mathrm{KL}}(P(x)\|Q(x)) \geq \mathbb{E}_{P(x)}[f(x)] - \log \mathbb{E}_{Q(x)}[\exp(f(x))], \tag{9}$$

where $f : \mathcal{X} \to \mathbb{R}$ is any measurable function such that $\mathbb{E}_{Q(x)}[\exp(f(x))] < \infty$.

B. Proposed Thompson Sampling Algorithm

The proposed algorithm described in Algorithm 1 implements two steps in each iteration $t \in \mathbb{N}$. The first step, called the *denoising step*, uses the current observed noisy context $\hat{c}_t$, and the history $\mathcal{H}_{t-1,c,\hat{c}} = \{c_\tau, \hat{c}_\tau\}_{\tau=1}^{t-1}$ of past observed noisy contexts and revealed true contexts, to obtain a *predictive posterior distribution* $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$ of the true context. This is a two-step process where firstly, we use the history $\mathcal{H}_{t-1,c,\hat{c}}$ of observations to update the agent's belief about the unknown noise channel parameter γ^* to a posterior distribution $P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})$. Thanks to the agent's knowledge of the prior $P(\gamma^*)$, the context distribution $P(c)$ as well as the noise channel likelihood $P(\hat{c}_t|c_t, \gamma^*)$, evaluating the posterior distribution is a consequence of applying the Baye's rule. The predictive posterior distribution is then obtained as $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathbb{E}_{P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})}[P(c_t|\hat{c}_t, \gamma^*)]$, where $P(c_t|\hat{c}_t, \gamma^*)$ is as defined in (2).

Algorithm 1: TS with Delayed Contexts ($\pi^{\mathrm{TS}}_{\mathrm{delay}}$)

1: Given parameters: $(\Sigma_n, \sigma^2, \lambda, \Sigma_\gamma, \mu_c, \Sigma_c)$. Initialize $\tilde{\mu}_0 = \mathbf{0} \in \mathbb{R}^m$ and $\tilde{\Sigma}_0^{-1} = (1/\lambda)\mathbb{I}$
2: **for** $t = 1, \ldots, T$ **do**
3: The environment selects a true context c_t.
4: Agent observes noisy context $\hat{c}_t$.
5: Agent computes $\tilde{R}_t$ and $\tilde{V}_t$ using (10) and (11) to evaluate $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathcal{N}(\tilde{V}_t, \tilde{R}_t^{-1})$.
6: Agent samples $\theta_t \sim \mathcal{N}(\tilde{\mu}_{t-1}, \tilde{\Sigma}_{t-1}^{-1})$ where $\tilde{\mu}_{t-1}$ and $\tilde{\Sigma}_{t-1}$ are defined as in (13) and (12).
7: Agent chooses action a_t as in (14).
8: Agent observes reward r_t corresponding to a_t, and the true context c_t.
9: **end for**

For linear-Gaussian bandits, the predictive posterior distribution is a multivariate Gaussian distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathcal{N}(\tilde{V}_t, \tilde{R}_t^{-1})$ with the inverse of covariance matrix obtained as

$$\tilde{R}_t = M - \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1} \tag{10}$$

and the mean vector obtained as

$$\tilde{V}_t = \tilde{R}_t^{-1}\Big(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}\hat{c}_t + \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1}\sum_{\tau=1}^{t-1}(\hat{c}_\tau - c_\tau) - \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1}M^{-1}(\Sigma_c^{-1}\mu_c - \Sigma_n^{-1}\hat{c}_t)\Big), \tag{11}$$

where $M = \Sigma_c^{-1} + \Sigma_n^{-1}$ and $\tilde{H}_t = \Sigma_n^{-1}M^{-1}\Sigma_n^{-1} + (t-1)\Sigma_n^{-1} + \Sigma_\gamma^{-1}$.

In the second step, we implement conventional Thompson sampling where we use the posterior distribution $P(\theta^*|\mathcal{H}_{t-1,r,a,c})$ of the reward parameter θ^* to sample $\theta_t \sim P(\theta^*|\mathcal{H}_{t-1,r,a,c})$. For the Gaussian bandit with Gaussian prior on θ^*, the posterior distribution $P(\theta^*|\mathcal{H}_{t-1,r,a,c}) = \mathcal{N}(\tilde{\mu}_{t-1}, \tilde{\Sigma}_{t-1}^{-1})$ is a multivariate Gaussian whose mean $\tilde{\mu}_{t-1}$ and variance $\tilde{\Sigma}_{t-1}^{-1}$ are determined by the observed history $\mathcal{H}_{t-1,r,a,c}$ as

$$\tilde{\Sigma}_{t-1} = \frac{1}{\lambda}\mathbb{I} + \frac{1}{\sigma^2}\sum_{\tau=1}^{t-1}\phi(a_\tau, c_\tau)\phi(a_\tau, c_\tau)^\top \tag{12}$$

$$\tilde{\mu}_{t-1} = \frac{\tilde{\Sigma}_{t-1}^{-1}}{\sigma^2}\Big(\sum_{\tau=1}^{t-1} r_\tau\phi(a_\tau, c_\tau)\Big). \tag{13}$$

Using the sampled θ_t and the obtained predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$, the agent then chooses action a_t as

$$a_t = \arg\max_{a\in\mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta_t, \tag{14}$$

where we have defined

$$\psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}}) := \mathbb{E}_{P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})}[\phi(a_t, c_t)] \tag{15}$$

as the expected feature map with respect to the inferred predictive distribution.

C. Bayesian Cumulative Regret Analysis

In this section, we focus on deriving an upper bound on the Bayesian cumulative regret, defined in (7), for the proposed TS algorithm. To this end, we define

$$\hat{a}_t = \arg\max_{a\in\mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^* \tag{16}$$

as the optimal action maximizing the mean reward $\psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*$ corresponding to the reward parameter θ^*. This can be interpreted as the optimal action taken by the agent had it known the reward parameter θ^*.

The Bayesian cumulative regret (7) of Algorithm 1 ($\pi^{\rm TS}_{\rm delay}$) can be then decomposed into three terms by adding and subtracting the term $\psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^* - \psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*$ as

$$\mathcal{R}^T(\pi^{\rm TS}_{\rm delay}) = \mathcal{R}^T_{\rm d,CB} + \mathcal{R}^T_{\rm d,EE1} + \mathcal{R}^T_{\rm d,EE2} \text{ where,} \tag{17}$$

$$\mathcal{R}^T_{\rm d,CB} = \sum_{t=1}^T \mathbb{E}\Big[\psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^* - \psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*\Big],$$

$$\mathcal{R}^T_{\rm d,EE1} = \sum_{t=1}^T \mathbb{E}\Big[\psi(a^*_t, \hat{c}_t|\gamma^*)^\top\theta^* - \psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*\Big],$$

$$\mathcal{R}^T_{\rm d,EE2} = \sum_{t=1}^T \mathbb{E}\Big[\psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^* - \psi(a_t, \hat{c}_t|\gamma^*)^\top\theta^*\Big].$$

Consequently, an upper bound on $\mathcal{R}^T(\pi^{\rm TS}_{\rm delay})$ follows by separately bounding each of the three terms in (17).

In (17), the first term $\mathcal{R}^T_{\rm d,CB}$ corresponds to the Bayesian cumulative regret of a standard contextual bandit with mean reward function $\psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*$ for $a \in \mathcal{A}$. We derive an upper bound on this term via an information-theoretic analysis motivated by the approach of [16]. To this end, we first make the following assumption.

Assumption 3.1: The feature map $\phi(\cdot,\cdot) \in \mathbb{R}^m$ has bounded norm, i.e., $\|\phi(\cdot,\cdot)\|_2 \le 1$.

The following lemma presents an upper bound on $\mathcal{R}^T_{\rm d,CB}$.

Lemma 3.1: Under Assumption 3.1, the following upper bound on $\mathcal{R}^T_{\rm d,CB}$ holds for $\frac{\lambda}{\sigma^2} \le 1$,

$$\mathcal{R}^T_{\rm d,CB} \le U_{\rm CB}(m,\lambda) := \sqrt{2Tm\sigma^2 \min\{m, 2(1+\log K)\}\log\Big(1+\frac{T\lambda}{m\sigma^2}\Big)}. \tag{18}$$

Proof: The proof follows from [16, Lemma 7] and [16, Lemma 3] and we skip the proof here. ■

The second term $\mathcal{R}^T_{\rm d,EE1}$in (17) accounts for the average difference in respective cumulative mean rewards of the oracle optimal action policy (4), that uses the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$, and the policy (16), that uses the inferred predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$. Thus, $\mathcal{R}^T_{\rm d,EE1}$ captures the error in approximating the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$ via the inferred predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})$. We show in the following lemma that the above approximation error over T iterations can be quantified, on average, via the mutual information $I(\gamma^*; \mathcal{H}_{T,c,\hat{c}})$ between γ^* and the T-length history of observed true and noisy contexts. The bound above also holds for the third term $\mathcal{R}^T_{\rm d,EE2}$ of (17) which similarly accounts for the average approximation error.

The following lemma thus presents an upper bound on the sum $\mathcal{R}^T_{\rm d,EE1} + \mathcal{R}^T_{\rm d,EE2}$.

Lemma 3.2: Under Assumption 3.1, for any $\delta \in (0,1)$, we have the following upper bound,

$$\mathcal{R}^T_{\rm d,EE1} + \mathcal{R}^T_{\rm d,EE2} \le 2\mathcal{R}^T_{\rm d,EE1} \le 4\sqrt{m\lambda T\log\Big(\frac{2m}{\delta}\Big) I(\gamma^*; \mathcal{H}_{T,c,\hat{c}})} + 2T\delta^2\sqrt{\frac{2m\lambda}{\pi}}. \tag{19}$$

Furthermore, if $\Sigma_\gamma = \sigma^2_\gamma\mathbb{I}$ and $\Sigma_n = \sigma^2_n\mathbb{I}$ for $\sigma^2_\gamma, \sigma^2_n > 0$, we have

$$I(\gamma^*; \mathcal{H}_{T,c,\hat{c}}) = \frac{d}{2}\log\Big(1 + (T-1)\frac{\sigma^2_\gamma}{\sigma^2_n}\Big).$$

Proof: See Appendix I. ■

Combining Lemma 3.1 and Lemma 3.2 then gives us the upper bound on $\mathcal{R}^T(\pi^{\rm TS}_{\rm delay})$.

Theorem 3.1: Under the setting of Lemma 3.1 and Lemma 3.2, the following inequality holds for $\delta \in (0,1)$

$$\mathcal{R}^T(\pi^{\rm TS}_{\rm delay}) \le U_{\rm CB}(m,\lambda) + 2T\delta^2\sqrt{\frac{2m\lambda}{\pi}} + 2\sqrt{2\lambda mTd\log\Big(1+(T-1)\frac{\sigma^2_\gamma}{\sigma^2_n}\Big)\log\Big(\frac{2m}{\delta}\Big)}. \tag{20}$$

Theorem 3.1 shows that Algorithm 1 achieves $\tilde{O}(d\sqrt{T})$[1] regret with the choice of $\delta = 1/T$ if $m > 2(1+\log K)$ and assuming m/d to be a constant.

IV. Experiments

In this section, we validate the performance of our algorithm via experiments on synthetic datasets. We study a data compression application in which the action involves selecting the compression matrix for the data/signal (context vector). We consider d-dimensional context vectors generated according to a Gaussian context distribution with mean zero and identity covariance matrix. The action set $\mathcal{A} = \{A_a\}_{a=1,\ldots,K}$ comprises of $K = 20$, $m \times d$ dimensional compression matrices A_a. The mean reward function corresponding to ath action and context vector c is obtained as $\phi(a,c)^\top\theta^* = (A_a c)^\top\theta^*$. We fix $\sigma^2 = 2$, $\lambda = 0.01$, $\Sigma_\gamma = \sigma^2_\gamma\mathbb{I}$, $\Sigma_n = \sigma^2_n\mathbb{I}$ for $\sigma^2_\gamma = \sigma^2_n = 3$.

We compare our algorithm with two baselines: TS_naive and TS_oracle. In TS_naive, the agent observes only noisy contexts but is unaware of the presence of noise. Consequently, it naively implements conventional TS with noisy context $\hat{c}_t$. This sets the benchmark for the worst-case achievable regret. The second baseline TS_oracle assumes that the agent knows the true channel parameter γ^*, a setting studied in [15], and can thus perform exact denoising via the predictive posterior

[1] $\tilde{O}(\cdot)$ omits logarithmic factors.

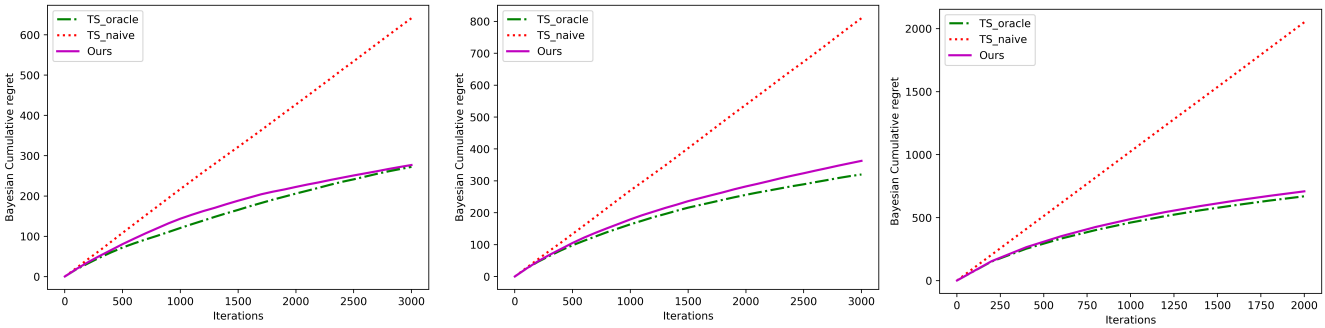

Fig. 1: Comparison of Bayesian regret of our algorithm with the baselines as a function of number of iterations: *(Left)* $d = 12$, $m = 6$; *(Center)* $d = 12$, $m = 10$; *(Right)* $d = 50$, $m = 30$. Other parameters are set as $K = 20$, $\sigma_n^2 = \sigma_\gamma^2 = 3$, $\lambda = 0.01$, $\sigma^2 = 2$. We run the algorithm for 10 independent trials for (Left) and (Center) figures, and for 100 trials for the (Right) figure, and plot the mean in the figures.

distribution $P(c_t|\hat{c}, \gamma^*)$. This algorithm sets the benchmark for the best achievable regret.

Fig. 1(Left) compares the Bayesian cumulative regret of our algorithm against the baselines when the matrix A_a compresses the 12-dimensional context vector to an $m = 6$ dimensional vector; Fig. 1(Center) compares them when m is increased to 10; and Fig. 1(Right) compares them when $d = 50$ and $m = 30$. The numerical results corroborate our theoretical findings: Our algorithm demonstrates sub-linear regret and achieve robust performance comparable to the best achievable performance of TS_oracle.

V. Conclusion and Future Work

We studied a stochastic CB problem where the agent observes noisy contexts through a noise channel with unknown channel parameter. For Gaussian bandits and Gaussian context noise, we introduced a TS algorithm that achieves $\tilde{O}(d\sqrt{T})$ Bayesian regret. We believe that the algorithm and key lemmas can be extended to when the likelihood-prior form conjugate distributions. Extension to general distributions is left for future work.

Appendix I
Proof of Lemma 3.2

Before stating the proof, we introduce some preliminary definitions and results.

Definition 1.1 (Sub-Gaussian Random Variable): A random variable y is said to be s^2-sub-Gaussian with respect to the distribution $P(y)$ if the following inequality holds:

$$\mathbb{E}_{P(y)}[\exp(\lambda(y - \mathbb{E}_{P(y)}[y]))] \leq \exp\Big(\frac{\lambda^2 s^2}{2}\Big).$$

Lemma 1.1 (Change of Measure Inequality): Let $x \in \mathbb{R}^n$ be a random vector and $g : \mathbb{R}^n \to \mathbb{R}$ denote a real-valued function. Let $P(x)$ and $Q(x)$ be two probability distributions defined on the space of x. If $g(x)$ is s^2-sub-Gaussian with respect to $Q(x)$, then the following inequality holds,

$$|\mathbb{E}_{P(x)}[g(x)] - \mathbb{E}_{Q(x)}[g(x)]| \leq \sqrt{2s^2 D_{\mathrm{KL}}(P(x)\|Q(x))}. \tag{21}$$

Proof: The inequality (21) follows by using the Donsker-Varadhan inequality (9) with $f(x) = \lambda g(x)$ for $\lambda \in \mathbb{R}$, using the sub-Gaussianity of $g(x)$ and finally optimizing over λ. ■

We are now read to prove Lemma 3.2. To prove an upper bound on the term $\mathcal{R}^T_{\mathrm{d,EE1}}$, we start by defining the following event:

$$\mathcal{E} := \Big\{\|\theta^*\|_2 \leq U\Big\}, \quad U := \sqrt{2\lambda m \log\Big(\frac{2m}{\delta}\Big)}. \tag{22}$$

Note that since $\theta^* \sim \mathcal{N}(\theta^*|\mathbf{0}, \lambda\mathbb{I})$, with probability at least $1 - \delta$, the following inequality holds $\|\theta^*\|_\infty \leq \sqrt{2\lambda \log\Big(\frac{2m}{\delta}\Big)} := u$. Since $\|\theta^*\|_2 \leq \sqrt{m}\|\theta^*\|_\infty$, the above inequality gives that the probability of the event $P(\mathcal{E}) \geq 1 - \delta$. We then have the following series of relations where we use $\Delta\Big(P(c_t|\hat{c}_t, \gamma^*), P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})\Big) = \mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t^*, c_t)^\top\theta^*] - \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})}[\phi(a_t^*, c_t)^\top\theta^*]$ as the difference in expectations of $\phi(a_t^*, c_t)^\top\theta^*$ with respect to $P(c_t|\hat{c}_t, \gamma^*)$ and $P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})$. Specifically, we get that

$$\begin{aligned}
\mathcal{R}^T_{\mathrm{d,EE1}} &\overset{(a)}{\leq} \sum_{t=1}^T \mathbb{E}\Big[\psi(a_t^*, \hat{c}_t|\gamma^*)^\top\theta^* - \psi(a_t^*, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*\Big] \\
&= \sum_{t=1}^T \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t, \gamma^*), P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})\Big)\Big] \\
&\overset{(b)}{\leq} \sum_{t=1}^T \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t, \gamma^*), P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big] \\
&\quad + 2\delta T\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c],
\end{aligned} \tag{23}$$

where the inequality (a) follows from the definition of $\hat{a}_t = \arg\max_{a\in\mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*$, and $\mathbf{1}\{\bullet\}$ denotes the indicator function that takes value 1 when $\bullet$ is true and is 0 otherwise. The inequality in (b) follows by noting that

$$\begin{aligned}
&\mathbb{E}[\Delta\Big(P(c_t|\hat{c}_t, \gamma^*), P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})\Big)\mathbb{I}\{\mathcal{E}^c\}] \\
&\leq \mathbb{E}\Big[\|\mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t^*, c_t)] - \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})}[\phi(a_t^*, c_t)]\|_2 \\
&\qquad \times \|\theta^*\|_2\mathbf{1}\{\mathcal{E}^c\}\Big] \\
&\leq 2\mathbb{E}\Big[\|\theta^*\|_2\mathbf{1}\{\mathcal{E}^c\}\Big] = 2P(\mathcal{E}^c)\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c] \leq 2\delta\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c],
\end{aligned} \tag{24}$$

where the last inequality is due to $P(\mathcal{E}^c) = 1 - P(\mathcal{E}) \leq \delta$. To obtain an upper bound on $\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c]$, we note that the following set of inequalities hold:

$$\begin{aligned}
\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c] &\overset{(a)}{\leq} \sqrt{m}\mathbb{E}[\|\theta^*\|_\infty|\mathcal{E}^c] \overset{(b)}{=} \sqrt{m}\mathbb{E}[\|\theta^*\|_\infty|\|\theta^*\|_\infty > u] \\
&= \sqrt{m}\sum_{i=1}^{m} P(\|\theta^*\|_\infty = |\theta_i^*|)\mathbb{E}\Big[\|\theta^*\|_\infty\Big|u < \|\theta^*\|_\infty = |\theta_i^*|\Big] \\
&\leq \sqrt{m}\sum_{i=1}^{m} \mathbb{E}\Big[|\theta_i^*|\Big||\theta_i^*| > u\Big] \\
&\overset{(c)}{=} 2\sqrt{m}\sum_{i=1}^{m}\int_{x>u} xg(x)dx \overset{(d)}{=} -2\lambda\sqrt{m}\sum_{i=1}^{m}\int_{x>u} g'(x)dx \\
&= 2\lambda m^{3/2} g(u) = 2\lambda m^{3/2}\frac{1}{\sqrt{2\pi\lambda}}\exp(-u^2/2\lambda) \\
&= \delta\sqrt{\frac{m\lambda}{2\pi}}, \qquad (25)
\end{aligned}$$

where (a) follows since $\|\theta\|_2 \leq \sqrt{m}\|\theta\|_\infty$, (b) follows since $\|\theta^*\|_2 > \sqrt{m}\sqrt{2\lambda\log\Big(\frac{2m}{\delta}\Big)}$ implies that $\|\theta^*\|_\infty > \sqrt{2\lambda\log\Big(\frac{2m}{\delta}\Big)} := u$. The equality in (c) follows by noting that $|\theta_i^*|$, where $\theta_i^* \sim \mathcal{N}(0,\lambda)$, follows a folded Gaussian distribution with density $2g(\theta_i^*)$ with $g(\theta_i^*) = \frac{1}{\sqrt{2\pi\lambda}}\exp(-\theta_i^{*2}/(2\lambda))$ being the Gaussian density. The equality in (d) follows by noting that $xg(x) = -\lambda g'(x)$, where $g'(x)$ is the derivative of the Gaussian density.

Furthermore, to obtain an upper bound on $\sum_{t=1}^{T}\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big]$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big] \\
&\leq P(\mathcal{E})\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)|\mathcal{E}\Big] \\
&\leq \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)|\mathcal{E}\Big]. \qquad (26)
\end{aligned}$$

Note that under the event $\mathcal{E}$, we have the following relation, $|\phi(a_t^*,c_t)^\top\theta^*)| \leq \|\phi(a_t^*,c_t)\|_2\|\theta^*\|_2 \leq U$, whereby $\phi(a_t^*,c_t)^\top\theta^*$ is U^2-sub-Gaussian. Consequently, applying Lemma 1.1 gives the following upper bound

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\Big[\sqrt{2U^2 D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}}))}\Big] \\
&\leq \sqrt{2TU^2\sum_{t=1}^{T}\mathbb{E}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}}))\Big]} \\
&\overset{(a)}{=} \sqrt{2TU^2\sum_{t=1}^{T} I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})} \\
&\overset{(b)}{\leq} \sqrt{2TU^2\sum_{t=1}^{T} I(c_t,\hat{c}_t;\gamma^*|\mathcal{H}_{c,\hat{c}})} \qquad (27) \\
&\overset{(c)}{=} \sqrt{2TU^2 I(\mathcal{H}_{T,c,\hat{c}};\gamma^*)} \qquad (28)
\end{aligned}$$

where the second inequality follows from applying Jensen's inequality; (a) follows by the definition of condition mutual information; (b) follows since $I(c_t,\hat{c}_t;\gamma^*|\mathcal{H}_{c,\hat{c}}) \geq I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})$ due to the non-negativity of mutual information; and finally, (c) follows from the chain rule of mutual information.

Using (28) and (25) in (23) gives us the bound in (19). Furthermore, analyzing the mutual information $I(\gamma^*;\mathcal{H}_{T,c,\hat{c}})$ for $\Sigma_\gamma = \sigma_\gamma^2\mathbb{I}$ and $\Sigma_n = \sigma_n^2\mathbb{I}$ yields $I(\gamma^*;\mathcal{H}_{T,c,\hat{c}}) = \frac{1}{2}d\log\Big(1+(T-1)\frac{\sigma_\gamma^2}{\sigma_n^2}\Big)$. Finally, note that the same upper bound holds for the term $\mathcal{R}_{\mathrm{d,EE2}}^T$.

REFERENCES

[1] T. Nakamura, N. Hayashi, and M. Inuiguchi, "Cooperative learning for adversarial multi-armed bandit on open multi-agent systems," *IEEE Control Systems Letters*, 2023.

[2] V. Srivastava, P. Reverdy, and N. E. Leonard, "Surveillance in an abruptly changing world via multiarmed bandits," in *IEEE Conference on Decision and Control (CDC)*, 2014, pp. 692–697.

[3] M. Aziz, E. Kaufmann, and M.-K. Riviere, "On multi-armed bandit designs for dose-finding clinical trials," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 686–723, 2021.

[4] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[5] X. Chen, Y. Nie, and N. Li, "Online residential demand response via contextual multi-armed bandits," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 433–438, 2020.

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

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[8] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 208–214.

[9] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, 2013, pp. 127–135.

[10] S. Lamprier, T. Gisselbrecht, and P. Gallinari, "Profile-based bandit with unknown profiles," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2060–2099, 2018.

[11] J. Kirschner and A. Krause, "Stochastic bandits with context distributions," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14 113–14 122, 2019.

[12] L. Yang, J. Yang, and S. Ren, "Multi-feedback bandit learning with probabilistic contexts," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main track*, 2020.

[13] J.-h. Kim, S.-Y. Yun, M. Jeong, J. Nam, J. Shin, and R. Combes, "Contextual linear bandits under noisy features: Towards bayesian oracles," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 1624–1645.

[14] S. T. Jose and S. Moothedath, "Thompson sampling for stochastic bandits with noisy contexts: An information-theoretic regret analysis," *arXiv:2401.11565*, 2024.

[15] H. Park and M. K. S. Faradonbeh, "Analysis of thompson sampling for partially observable contextual multi-armed bandits," *IEEE Control Systems Letters*, vol. 6, pp. 2150–2155, 2021.

[16] G. Neu, I. Olkhovskaia, M. Papini, and L. Schwartz, "Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9486–9498, 2022.