entropy

MDPI

Article

Thompson Sampling for Stochastic Bandits with Noisy Contexts: An Information-Theoretic Regret Analysis

Sharu Theresa Jose [1,*] and Shana Moothedath [2,*]

[1] School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
[2] Department of Electrical Engineering, Iowa State University, Ames, IA 50011, USA
* Correspondence: s.t.jose@bham.ac.uk (S.T.J.); mshana@iastate.edu (S.M.)

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 parameters. 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 Thompson sampling algorithm and analyze its Bayesian cumulative regret with respect to the oracle action policy via information-theoretic tools. For Gaussian bandits with Gaussian context noise, our information-theoretic analysis shows that under certain conditions on the prior variance, the Bayesian cumulative regret scales as $\tilde{O}(m\sqrt{T})$, where m is the dimension of the feature vector and T is the time horizon. We also consider the problem setting where the agent observes the true context with some delay after receiving the reward, and show that delayed true contexts lead to lower regret. Finally, we empirically demonstrate the performance of the proposed algorithms against baselines.

Keywords: noisy contextual bandits; Thompson sampling; Bayes regret; information theory

Citation: Jose, S.T.; Moothedath, S. Thompson Sampling for Stochastic Bandits with Noisy Contexts: An Information-Theoretic Regret Analysis. *Entropy* **2024**, *26*, 606. https://doi.org/10.3390/e26070606

Academic Editors: Christos Makris, Vasileios Megalooikonomou, Sotiris Kotsiantis and Isidoros Perikos

Received: 20 May 2024
Revised: 9 July 2024
Accepted: 15 July 2024
Published: 17 July 2024

1. Introduction

Decision-making in the face of uncertainty is a widespread challenge found across various domains such as control and robotics [1], clinical trials [2], communications [3], and ecology [4]. To tackle this challenge, learning algorithms have been developed to uncover effective policies for optimal decision-making. One notable framework for addressing this is contextual bandits (CBs), which capture the essence of sequential decision-making by incorporating side information, termed *context* [5].

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 reward is stochastic, drawn from a probability distribution whose mean reward (which is a function of context-action pair) is unknown to the agent. 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.

In this paper, we focus on a CB model that assumes stochastic rewards with linear mean-reward functions, also called stochastic linear contextual bandits. Stochastic linear CB models find applications in various settings including internet advertisement selection [6], where the advertisement (i.e., action) and webpage features (i.e., context) are used to construct a linear predictor of the probability that a user clicks on a given advertisement, and article recommendation on web portals [7].

While most prior research on CBs has primarily focused on models with known exact contexts [8–10], in many real-world applications, the contexts are noisy, e.g., imprecise measurement of patient conditions in clinical trials, weather or stock market predictions. 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. Consequently, recent efforts have been made to develop CB algorithms tailored to noisy context settings.

Related Works: ref. [11] considers a setting where there is a bounded zero-mean noise in the m-dimensional *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. Ref. [12] 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. A similar setting has also been considered in [13]. Differing from these works, ref. [14] 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 Optimism in the Face of Uncertainty (OFUL) algorithm. A variant of this setting has been studied in [15].

Motivation and Problem Setting: In this work, inspired by [14], we consider the following noisy CB setting. In each round, the environment samples a true context vector c_t from a *context distribution* that is *known* to the agent. The agent, however, does not observe the true context but 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 is aware of the noise present but does not know the channel parameter γ^*. Following [14], we consider Gaussian noise channels for our regret analysis.

Based on the observed noisy contexts, the agent chooses an action a_t and observes a reward r_t corresponding to the true context. We consider a linear bandit whose mean reward $\phi(a_t, c_t)^\top \theta^*$ is determined by an unknown reward parameter θ^*. 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.

Our setting differs from [14] in that we assume noisy contexts rather than noisy feature vectors and that the agent knows the context distribution. The noise model, incorporating noise in the feature vector, allows [14] to transform the original problem into a different CB problem that estimates a modified reward parameter. Such a transformation, however, is not straightforward in our setting with noise in contexts rather than in feature vectors, where we wish to analyze the Bayesian regret. Additionally, we propose a de-noising approach to estimate the predictive distribution of the true context from given noisy contexts, offering potential benefits for future analyses.

The assumption of known context distribution follows from [12]. This can be motivated by considering the example of an online recommendation engine that pre-processes the user account registration information or contexts (e.g., age, gender, device, location, item preferences) to group them into different clusters [16]. The engine can then infer the 'empirical' distribution of users within each cluster to define a context distribution over true contextual information. A noisy contextual information scenario occurs when a guest with different preferences logs into a user's account.

Challenges and Novelty: Different from existing works that developed UCB-based algorithms, we propose a fully Bayesian Thompson Sampling (TS) algorithm that approximates the Bayesian oracle policy. The proposed algorithm differs from the standard contextual TS [10] in the following aspects. Firstly, since the true context vectors are not accessible at each round and the channel parameter γ^* is unknown, the agent uses its knowledge of the context distribution and the past observed noisy 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*. Secondly, the reward r_t received by the agent corresponds to the unobserved true context c_t. Hence, the agent cannot accurately evaluate

the posterior distribution of θ^* and sample from it as is conducted in standard contextual TS. Instead, our algorithm proposes to use a sampling distribution that 'approximates' the posterior.

Different from existing works that focus on frequentist regret analysis, we derive novel *information-theoretic* bounds on the *Bayesian cumulative regret* of our algorithm. For Gaussian bandits, our information-theoretic regret bounds scale as $\tilde{O}(m\sqrt{T})$ (the notation $\tilde{O}(\bullet)$ suppresses logarithmic terms in $\bullet$), where m, T denote the dimension of the feature vector and time horizon respectively, under certain conditions on the variance of the prior on θ^*. Furthermore, our Bayesian regret analysis shows that the *posterior mismatch*, resulting due to replacing the true posterior distribution with a sampling distribution, results in an approximation error that is captured via the Kullback–Leibler (KL) divergence between the distributions. To the best of our knowledge, quantifying the posterior mismatch via KL divergence has not been studied before and is of independent interest.

Finally, we also extend our algorithm to a setting where the agent observes the true context after the decision is made and reward is observed [12]. We call this setting CBs with delayed true contexts. Such scenarios arise in many applications where only a prediction of the context is available at the time of decision-making; however, the true context is available later. For instance, in farming-recommender systems where, at the time of making the decision regarding which crop to cultivate in a year, the true contextual information about the weather pattern is unavailable, while some 'noisy' weather predictions are available. In fact, the true weather pattern is observed only after the decision is made. We show that our TS algorithm for this setting with delayed true contexts results in reduced Bayesian regret. Table 1 compares our regret bound with that of the state-of-the-art algorithms in the noiseless and noisy CB settings.

Table 1. Comparison of the regret bounds of our proposed TS algorithm for noisy CB with state-of-the art algorithms.

Reference	Setting	Algorithm	Regret	Bound
[8]	Linear CB	LinRel	Frequentist	$\tilde{O}(\sqrt{mT})$
[9]	Linear CB	Lin-UCB	Frequentist	$\tilde{O}(\sqrt{mT})$
[10]	Linear CB	TS	Frequentist	$O(m\sqrt{T}\log^{3/2} T)$
[17]	Linear CB	TS	Bayesian	$O(m\sqrt{T}\log T)$
[11]	Noisy CB	SampLinUCB	Frequentist	$\tilde{O}(m\sqrt{T})$
[12]	Noisy CB	UCB	Frequentist	$\tilde{O}(m\sqrt{T})$
[14]	Noisy CB	OFUL	Frequentist	$\tilde{O}(m\sqrt{T})$
Our work	**Noisy CB**	**TS**	**Bayesian**	$\tilde{O}(m\sqrt{T})$

2. 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. At iteration $t \in \mathbb{N}$, the environment randomly draws a context vector $c_t \in \mathcal{C}$ according to a *context distribution* $P(c)$ defined over the space $\mathcal{C}$ of context vectors. The context distribution $P(c)$ is known to the agent. The agent, however, does not observe the true context c_t drawn by the environment. Instead, it observes a noisy version $\hat{c}_t$ of the true context, 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. The noise channel $P(\hat{c}_t|c_t, \gamma^*)$ is parameterized by the *noise channel parameter* γ^* that is *unknown* to the agent.

Having observed the noisy context $\hat{c}_t$ at iteration t, the agent chooses an action $a_t \in \mathcal{A}$ according to an *action policy* $\pi_t(\cdot|\hat{c}_t)$. The action policy may be stochastic describing a

probability distribution over the set $\mathcal{A}$ of actions. Corresponding to the chosen action a_t, the agent receives a reward from the environment given by

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

where $f(\theta^*, a_t, c_t) = \phi(a_t, c_t)^\top \theta^*$ is the linear *mean-reward function* and ξ_t is a zero-mean reward noise variable. The mean reward function $f(\theta^*, a_t, c_t)$ is defined via the *feature map* $\phi : \mathcal{A} \times \mathcal{C} \to \mathbb{R}^m$, that maps the action and true context to an m-dimensional feature vector, and via the reward parameter $\theta^* \in \mathbb{R}^m$ that is *unknown* to the agent.

We call the noisy CB problem described above *CBs with unobserved true context* (see Setting 1) since the agent does not observe the true context c_t and the selection of action is based solely on the observed noisy context. Accordingly, at the end of iteration t, the agent has accrued the history $\mathcal{H}_{t,r,a,\hat{c}} = \{r_\tau, a_\tau, \hat{c}_\tau\}_{\tau=1}^t$ of observed reward-action-noisy context tuples. The action policy $\pi_{t+1}(\cdot|\hat{c}_{t+1})$ at $(t+1)^{\text{th}}$ iteration may depend on the history $\mathcal{H}_{t,r,a,\hat{c}}$.

Setting 1: CBs with unobserved true contexts

1: **for** $t = 1, \ldots, T$ **do**
2: Environment samples $c_t \sim P(c)$.
3: Agent observes noisy context $\hat{c}_t \sim P(\hat{c}_t|c_t, \gamma^*)$.
4: Agent chooses an action $a_t \sim \pi_t(\cdot|\hat{c}_t)$.
5: Agent receives reward r_t according to (1).
6: **end for**

We also consider a variant of the above problem where the agent has access to a *delayed* observation of the true context c_t as studied in [12]. We call this setting *CBs with delayed true context*. In this setting, at iteration t, the agent first observes a noisy context $\hat{c}_t$, chooses action $a_t \sim \pi_t(\cdot|\hat{c}_t)$, and receives reward r_t. Later, the true context c_t is observed. It is important to note that the agent has no access to the true context at the time of decision-making. Thus, at the end of iteration t, the agent has collected 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.

In both of the problem settings described above, the agent's objective 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.

Bayesian Cumulative Regret

The cumulative regret of an action policy $\pi_t(\cdot|\hat{c}_t)$ quantifies how different 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 $\psi(a,\hat{c}|\gamma^*) := \mathbb{E}_{P(c|\hat{c},\gamma^*)}[\phi(a,c)]$. Note, that as in [14,18], 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 γ^*.

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

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

the expected difference in mean rewards of the oracle decision policy and the agent's decision policy over T iterations. In (5), the expectation is taken with respect to the joint distribution $P(\mathcal{H}_{t-1,r,\hat{c},c,a}|\theta^*,\gamma^*)P(\hat{c}_t,c_t,a_t|\mathcal{H}_{t-1,r,\hat{c},c,a},\theta^*,\gamma^*)$, where $P(\hat{c}_t,c_t,a_t|\mathcal{H}_{t-1,r,\hat{c},c,a},\theta^*,\gamma^*) = P(\hat{c}_t,c_t|\gamma^*)\pi_t(a_t|\hat{c}_t) = P(\hat{c}_t|\gamma^*)P(c_t|\hat{c}_t,\gamma^*)\pi_t(a_t|\hat{c}_t)$. Using this, the cumulative regret (5) can be written as

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

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 observe 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 priors $P(\theta^*)$ and $P(\gamma^*)$.

In the next sections, we present our novel TS algorithms to minimize the Bayesian cumulative regret for the two problem settings considered in this paper.

3. Modified TS for CB with Unobserved True Contexts

In this section, we consider Setting 1 where the agent only observes the noisy context $\hat{c}_t$ at each iteration t. Our proposed modified TS Algorithm is given in Algorithm 1.

Algorithm 1: TS with unobserved true contexts (π^{TS})

for $t = 1, \ldots, T$ **do**
  The environment selects a true context c_t.
  Agent observes noisy context $\hat{c}_t$.
  Agent evaluates the predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$ as in (8).
  Agent samples $\theta_t \sim \bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$.
  Agent chooses action a_t as in (11).
  Agent observes reward r_t as in (1).
end for

The proposed algorithm implements two steps in each iteration $t \in \mathbb{N}$. In the first step, called the *de-noising* step, the agent uses the current observed noisy context $\hat{c}_t$ and the history $\mathcal{H}_{t-1,\hat{c}} = \{\hat{c}_\tau\}_{\tau=1}^{t-1}$ of past observed noisy contexts to obtain a *predictive posterior distribution* $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$ of the true context c_t. This is a two-step process, where firstly the agent uses the history $\mathcal{H}_{t-1,\hat{c}}$ of past observed noisy contexts to compute the posterior distribution of γ^* as $P(\gamma^*|\mathcal{H}_{t-1,\hat{c}}) \propto P(\gamma^*) \prod_{\tau=1}^{t-1} P(\hat{c}_\tau|\gamma^*)$, where the conditional distribution $P(\hat{c}_t|\gamma^*)$ is evaluated as in (3). Note, that to evaluate the posterior, the agent uses its knowledge of the context distribution $P(c)$, the prior $P(\gamma^*)$ and the noise channel likelihood $P(\hat{c}_t|c_t, \gamma^*)$. Using the derived posterior $P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})$, the predictive posterior distribution of the true context is then obtained as

$$P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}}) = \mathbb{E}_{P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})}[P(c_t|\hat{c}_t, \gamma^*)], \tag{8}$$

where $P(c_t|\hat{c}_t, \gamma^*)$ is defined as in (2).

The second step of the algorithm implements a *modified* Thompson sampling. Note, that since the agent does not have access to the true contexts, it cannot evaluate the posterior distribution with known contexts,

$$P(\theta^*|\mathcal{H}_{t-1,r,a,c}) \propto P(\theta^*) \prod_{\tau=1}^{t-1} P(r_\tau|a_\tau, c_\tau, \theta^*), \tag{9}$$

as is conducted in standard contextual TS. Instead, the agent must evaluate the true *posterior distribution* under noisy contexts,

$$\begin{aligned} P_t(\theta^*) &:= P(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}}) \\ &\propto P(\theta^*)\mathbb{E}_{P(\gamma^*)}\Big[\prod_{\tau=1}^{t-1} \mathbb{E}_{P(c_\tau)}[P(\hat{c}_\tau|c_\tau, \gamma^*)P(r_\tau|a_\tau, c_\tau, \theta^*)]\Big]. \end{aligned} \tag{10}$$

However, evaluating the marginal distribution $\mathbb{E}_{P(\gamma^*)}\Big[\prod_{\tau=1}^{t-1} \mathbb{E}_{P(c_\tau)}[P(\hat{c}_\tau|c_\tau, \gamma^*)P(r_\tau|a_\tau, c_\tau, \theta^*)]\Big]$ is challenging even for Gaussian bandits as the mean $\phi(a_\tau, c_\tau)^\top \theta^*$ of the reward distribution $P(r_\tau|a_\tau, c_\tau, \theta^*)$ is, in general, a non-linear function of the true context c_τ. As a result, the posterior $P_t(\theta^*)$ is analytically intractable.

Consequently, at each iteration t, the agent samples $\theta_t \sim \bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$ from a distribution $\bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$ that 'approximates' the true posterior $P_t(\theta^*)$. The specific choice of this sampling distribution depends on the problem setting. Ideally, one must choose a distribution that is sufficiently 'close' to the true posterior. In the next sub-section, we will explain the choice for Gaussian bandits.

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

$$a_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta_t, \text{where} \tag{11}$$

$$\psi(a_t, \hat{c}_t|\mathcal{H}_{\hat{c}}) := \mathbb{E}_{P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})}[\phi(a_t, c_t)] \tag{12}$$

is the expected feature map with respect to $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$.

3.1. Linear-Gaussian Stochastic CBs

We now instantiate Algorithm 1 for Gaussian CBs. Specifically, we consider Gaussian bandits with the reward noise ξ_t in (1) as Gaussian $\mathcal{N}(0, \sigma^2)$ with mean 0 and variance $\sigma^2 > 0$. We also assume a Gaussian prior $P(\theta^*) = \mathcal{N}(\mathbf{0}, \lambda\mathbb{I})$ on the reward parameter θ^* with mean zero and an $m \times m$ diagonal, covariance matrix with entries $\lambda > 0$. Here, $\mathbb{I}$ denotes the identity matrix. The assumption of diagonal prior covariance is in line with Lemma 3 in [19].

We consider a multivariate Gaussian context distribution $P(c) = \mathcal{N}(\mu_c, \Sigma_c)$ with mean $\mu_c \in \mathbb{R}^d$ and covariance matrix $\Sigma_c \in \mathbb{R}^{d\times d}$. The context noise channel $P(\hat{c}|c, \gamma^*)$ is also similarly Gaussian with a mean $(\gamma^* + c)$ and covariance matrix $\Sigma_n \in \mathbb{R}^{d\times d}$. We assume the prior on noise channel parameter γ^* to be Gaussian $P(\gamma^*) = \mathcal{N}(\mathbf{0}, \Sigma_\gamma)$ with d-dimensional zero mean vector $\mathbf{0}$ and covariance matrix $\Sigma_\gamma \in \mathbb{R}^{d\times d}$. We assume that Σ_c, Σ_γ and Σ_n are all positive definite matrices known to the agent. The assumption of positive definite covariance matrices is to facilitate the Bayesian analysis adopted in this work. Similar assumptions were also required in the related work of [14].

For this setting, we can analytically evaluate the predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}}) = \mathcal{N}(c_t|V_t, R_t^{-1})$ as a multi-variate Gaussian with inverse covariance matrix,

$$R_t = M - \Sigma_n^{-1}(H_t^{-1})^\top \Sigma_n^{-1}, \tag{13}$$

where $H_t = (t-1)\Sigma_n^{-1} - (t-2)\Sigma_n^{-1}M^{-1}\Sigma_n^{-1} + \Sigma_\gamma^{-1}$ and $M = \Sigma_c^{-1} + \Sigma_n^{-1}$, and with the mean vector

$$V_t = (R_t^{-1})^\top \Big(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}\hat{c}_t - \Sigma_n^{-1}(H_t^{-1})^\top L_t^\top\Big), \tag{14}$$

where

$$\begin{aligned} L_t^\top &= \Sigma_n^{-1}M^{-1}(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}\hat{c}_t) + (\Sigma_n^{-1} - \Sigma_n^{-1}M^{-1}\Sigma_n^{-1})\sum_{\tau=1}^{t-1}\hat{c}_\tau \\ &\quad - (t-1)\Sigma_n^{-1}M^{-1}\Sigma_c^{-1}\mu_c. \end{aligned}$$

Derivations are presented in Appendix C.1.2.

For the *modified*-TS step, we sample θ_t from the approximate posterior distribution

$$\bar{P}_t(\theta^*) := \bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}}) \propto P(\theta^*)\prod_{\tau=1}^{t-1}\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*), \tag{15}$$

where

$$\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*) = \mathcal{N}(\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})^\top\theta^*, \sigma^2) \tag{16}$$

and $\psi(a_t, \hat{c}_t|\mathcal{H}_{\hat{c}})$ is the expected feature map defined in (12). This yields the approximate posterior to be a Gaussian distribution $\bar{P}_t(\theta^*) = \mathcal{N}(\mu_{t-1}, \Sigma_{t-1}^{-1})$ whose inverse covariance matrix and mean, respectively, evaluate as

$$\Sigma_{t-1} = \frac{\mathbb{I}}{\lambda} + \frac{1}{\sigma^2}\sum_{\tau=1}^{t-1}\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})^\top \tag{17}$$

$$\mu_{t-1} = \frac{\Sigma_{t-1}^{-1}}{\sigma^2}\Big(\sum_{\tau=1}^{t-1} r_\tau\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})\Big). \tag{18}$$

The sampling distribution $\bar{P}_t(\theta^*)$ considered above is different from the true posterior distribution (10), which is analytically intractable. However, it bears resemblance to the

posterior (9) when the true contexts are known, with the reward distribution $P(r_\tau|a_\tau, c_\tau, \theta^*)$ replaced by $\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)$. In Section 3.2.2, we show that the above choice of sampling distribution is indeed 'close' to the true posterior.

3.2. Bayesian Regret Analysis

In this section, we derive information-theoretic upper bounds on the Bayesian regret (7) of the modified TS algorithm for Gaussian CBs. To this end, we first outline the key information-theoretic tools required to derive our bound.

3.2.1. Preliminaries

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

$$D_{\mathrm{KL}}(P(x)||Q(x)) = \mathbb{E}_{P(x)}\left[\log \frac{P(x)}{Q(x)}\right], \tag{19}$$

if $P(x)$ is absolutely continuous with respect to $Q(x)$, and takes value ∞ otherwise. If x and y denote two random variables described by the joint probability distribution $P(x, y)$, the mutual information $I(x; y)$ 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 evaluates as

$$I(x; y|z) = \mathbb{E}_{P(z)}[D_{KL}(P(x, y|z)\|P(x|z)P(y|z))]$$

where $P(x|z)$ and $P(y|z)$ are conditional distributions. We will also use the following variational representation of the KL-divergence, also termed the *Donskar–Varadhan* (*DV*) inequality,

$$D_{\mathrm{KL}}(P(x)\|Q(x)) \geq \mathbb{E}_{P(x)}[f(x)] - \log \mathbb{E}_{Q(x)}[\exp(f(x))], \tag{20}$$

which holds for any measurable function $f : \mathcal{X} \to \mathbb{R}$ satifying the inequality $\mathbb{E}_{Q(x)}[\exp(f(x))] < \infty$.

3.2.2. Information-Theoretic Bayesian Regret Bounds

In this section, we present information-theoretic upper bounds on the Bayesian regret of the modified TS algorithm. To this end, we first state our main assumption.

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

The following theorem gives our main result.

Theorem 1. *Assume that the covariance matrices satisfy $\Sigma_n\Sigma_c^{-1} \succ 0$ and $\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M \succ 0$ where $M = \Sigma_n^{-1} + \Sigma_c^{-1}$. Under Assumption 1, if $\frac{\lambda}{\sigma^2} \leq \frac{1}{T} \leq 1$, the following upper bound on the Bayesian regret of the modified TS algorithm holds,*

$$\mathcal{R}^T(\pi^{\mathrm{TS}}) \leq U(m, \frac{\sigma^2}{T}) + \sqrt{2Tm\sigma^2} + \sqrt{2T\sigma^2(\log(K) + m)}$$
$$+ \frac{4}{T}\sqrt{\frac{m\sigma^2}{2T\pi}} + 2\sqrt{4\sigma^2 m \log\Big(2mT\Big)\Big(\mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M)^{-1}) + \log(T)\mathrm{Tr}(\Sigma_c\Sigma_n^{-1})\Big)},$$

where

$$U(m, \lambda) = \sqrt{2Tm\sigma^2 \min\{m, 2\log(1+K)\}\log\Big(1 + \frac{T\lambda}{m\sigma^2}\Big)}. \tag{21}$$

The theorem above shows that the proposed TS algorithm achieves $\tilde{O}(m\sqrt{T})$ regret when the prior $P(\theta^*)$ is highly informative with variance parameter satisfying the constraint $\lambda \leq \sigma^2/T$.

Remark 1. *The assumption on covariance matrices in Theorem 1 directly holds for diagonal covariance matrices with positive eigen values.*

To prove the regret bound of Theorem 1, we start by defining

$$\hat{a}_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t | \mathcal{H}_{\hat{c}})^\top \theta^* \tag{22}$$

as the action that maximizes the mean reward $\psi(a, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^*$ corresponding to reward parameter θ^*. Using the above, the Bayesian cumulative regret (7) for the proposed TS algorithm π^{TS} can be decomposed as

$$\begin{aligned}
&\mathcal{R}^T(\pi^{\text{TS}}) = \mathcal{R}^T_{\text{CB}} + \mathcal{R}^T_{\text{EE1}} + \mathcal{R}^T_{\text{EE2}}, \text{where} \qquad (23)\\
&\mathcal{R}^T_{\text{CB}} = \sum_{t=1}^{T} \mathbb{E}\Big[\psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^* - \psi(a_t, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^*\Big],\\
&\mathcal{R}^T_{\text{EE1}} = \sum_{t=1}^{T} \mathbb{E}\Big[\psi(a^*_t, \hat{c}_t|\gamma^*)^\top \theta^* - \psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^*\Big],\\
&\mathcal{R}^T_{\text{EE2}} = \sum_{t=1}^{T} \mathbb{E}\Big[\psi(a_t, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^* - \psi(a_t, \hat{c}_t|\gamma^*)^\top \theta^*\Big].
\end{aligned}$$

In (23), the first term $\mathcal{R}^T_{\text{CB}}$ quantifies the Bayesian regret of our action policy (11) with respect to the action policy (22) for a CB with mean reward function $\psi(a, \hat{c}_t|\mathcal{H}_{\hat{c}})^\top \theta^*$. The second term $\mathcal{R}^T_{\text{EE1}}$ accounts for the average difference in the cumulative mean rewards of the oracle optimal action policy (4), evaluated using the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$, and our action policy (11), that uses the inferred predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$. In this sense, $\mathcal{R}^T_{\text{EE1}}$ captures the error in approximating the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$ via the inferred predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{\hat{c}})$. The third term $\mathcal{R}^T_{\text{EE2}}$ similarly accounts for the average approximation error.

To derive an upper bound on the Bayesian regret $\mathcal{R}^T(\pi^{\text{TS}})$, we separately upper bound each of the three terms in (23) as derived in the following lemmas. The lemma below presents an upper bound on $\mathcal{R}^T_{\text{CB}}$.

Lemma 1. *Under Assumption 1, the following upper bound holds if* $\frac{\lambda}{\sigma^2} \leq \frac{1}{T} \leq 1$,

$$\begin{aligned}
\mathcal{R}^T_{\text{CB}} &\leq U(m, \frac{\sigma^2}{T}) + \sqrt{2\sigma^2 \sum_{t=1}^{T} D_t} + \sqrt{2\sigma^2 \Big(T \log(K) + \sum_{t=1}^{T} D_t\Big)}, \qquad (24)\\
&\leq U(m, \frac{\sigma^2}{T}) + \sqrt{2Tm\sigma^2} + \sqrt{2T\sigma^2(\log(K) + m)}, \qquad (25)
\end{aligned}$$

where $D_t = \mathbb{E}[D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*))]$ *and* $U(m, \lambda)$ *is as defined in* (21).

To derive the upper bound in (24), we leverage results from [19] that study information-theoretic Bayesian regret of standard contextual TS algorithms via lifted information-ratio. However, the results do not directly apply to our algorithm due to the *posterior mismatch* between the sampling distribution $\bar{P}_t(\theta^*)$ and the true posterior distribution $P_t(\theta^*)$. Consequently, our upper bound (24) consists of three terms: the first term, defined as in (21), corresponds to the upper bound on the Bayesian regret of contextual TS that assumes $\bar{P}_t(\theta^*)$ as the true posterior. This can be obtained by applying the lifted information ratio-based analysis of Cor. 2 in [19]. The second and third terms account for the posterior mismatch

via the expected KL-divergence $\mathbb{E}[D_{\mathrm{KL}}(P_t(\theta^*)\|\bar{P}_t(\theta^*))]$ between the true posterior $P_t(\theta^*)$ and the sampling distribution $\bar{P}_t(\theta^*)$. In particular, this expected KL divergence can be upper bounded by $2(t-1)\lambda m/\sigma^2$ (See Appendix C.1.3 for proof) under the prior $P(\theta^*) = \mathcal{N}(\mathbf{0}, \lambda\mathbb{I})$. Importantly, our result holds when this prior distribution is sufficiently concentrated with its variance satisfying the inequality $\lambda \leq \sigma^2/T$. This ensures that the contribution of posterior mismatch to the Bayes regret scales is $O(\sqrt{mT})$.

The following lemma gives an upper bound on the sum $\mathcal{R}^T_{\mathrm{EE1}} + \mathcal{R}^T_{\mathrm{EE2}}$.

Lemma 2. *Under Assumption 1, the following upper bound holds for* $\delta \in (0,1)$,

$$\begin{aligned}\mathcal{R}^T_{\mathrm{EE1}} + \mathcal{R}^T_{\mathrm{EE2}} &\leq 2\mathcal{R}^T_{\mathrm{EE1}} \\ &\leq 4\delta^2 T\sqrt{\frac{m\lambda}{2\pi}} + 2\sqrt{4\lambda T m \log\Big(\frac{2m}{\delta}\Big)\sum_{t=1}^{T} I(\gamma^*; c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})}. \end{aligned} \tag{26}$$

In addition, if the covariance matrices satisfy that $\Sigma_n\Sigma_c^{-1} \succ 0$ *and* $\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M \succ 0$ *where* $M = \Sigma_n^{-1} + \Sigma_c^{-1}$, *then* (26) *can be further upper bounded as*

$$\begin{aligned}&\mathcal{R}^T_{\mathrm{EE1}} + \mathcal{R}^T_{\mathrm{EE2}} \\ &\leq 4\delta^2 T\sqrt{\frac{m\lambda}{2\pi}} + 2\sqrt{4\lambda T m \log\Big(\frac{2m}{\delta}\Big)\Big(\mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M)^{-1}) + \log(T)\mathrm{Tr}(\Sigma_c\Sigma_n^{-1})\Big)}. \end{aligned} \tag{27}$$

Lemma 2 shows that the error in approximating $P(c_t|\hat{c}_t, \gamma^*)$ with $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$, on average, can be quantified via the conditional mutual information $I(\gamma^*; c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$ between γ^* and true context c_t given knowledge of observed noisy contexts up to and including iteration t.

Finally, combining Lemmas 1 and 2 with the choice of $\delta = 1/T$ and $\lambda \leq \sigma^2/T$ gives us the regret bound in Theorem 1.

3.3. Beyond Gaussian Bandits

In the previous sections, we studied Gaussian bandits and analyzed the Bayesian regret. We will now discuss the potential extension of results beyond Gaussian bandits. As in [14], we will focus on Gaussian context distribution and context noise distribution, which helps to derive the upper bound on the estimation errors in Lemma 2.

To extend the Bayesian regret analysis to non-Gaussian bandits, Lemma 1 requires bandit-specific modifications. Specifically, the derivation of the term $U(m,\lambda)$, that captures the standard Bayesian regret of contextual TS with $\bar{P}_t(\theta^*)$ as the true posterior, and that of the posterior mismatch term via the expected KL divergence critically depends on the type of bandit and the choice of the sampling posterior. The Bayesian regret bound $U(m,\lambda)$ is derived using the lifted information ratio-based approach of [19]. This can indeed be extended to non-Gaussian bandits like logistic bandits (see [19]) to obtain a modified $U(m,\lambda)$ term.

However, the analysis of posterior mismatch term for non-Gaussian bandits is non-trivial and depends on the specific bandit assumed. Firstly, to characterize the posterior mismatch via the expected KL divergence, our analysis requires the chosen sampling distribution $\bar{P}_t(\theta^*)$ to be sub-Gaussian. To choose the sampling distribution, one can follow the framework adopted in (15) and (16) and use an 'appropriate' reward distribution $\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)$ such that (a) the KL divergence $D_{\mathrm{KL}}(P(r_\tau|a_\tau, c_\tau, \theta^*)\|\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*))$ between the true reward distribution and the chosen reward distribution is small to minimize posterior mismatch, and (b) the resulting sampling distribution is easy to sample from and has sub-Gaussian tails. Thus, analyzing the posterior mismatch for non-Gaussian bandits requires a case-by-case treatment. For Gaussian bandits, we control the above KL divergence by choosing a Gaussian distribution $\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)$ with mean $\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})$ as in (16). Finally, in Section 5, we

extend Algorithm 1 to logistic bandits with the choice of sampling distribution motivated by (15) and (16) and use Langevin Monte Carlo to sample from this distribution.

4. TS for CB with Delayed True Contexts

In this section, we consider the CBs with delayed true context setting where the agent observes the true context c_t after it observes the reward r_t corresponding to the chosen action a_t. Note, that at the time of choosing action a_t, the agent has access only to noisy contexts. We specialize our TS algorithm to this setting, and call it Algorithm 2 (or $\pi^{\text{TS}}_{\text{delay}}$).

Algorithm 2: TS for Delayed True contexts ($\pi^{\text{TS}}_{\text{delay}}$)

1: **for** $t = 1, \ldots, T$ **do**
2: The environment selects a true context c_t.
3: Agent observes noisy context $\hat{c}_t$.
4: Agent evaluates the predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$ as in (28).
5: Agent samples $\theta_t \sim P(\theta^*|\mathcal{H}_{t-1,r,a,c})$.
6: Agent chooses action a_t as in (33).
7: Agent observes reward r_t (as in (1)) and the true context c_t.
8: **end for**

Algorithm 2 follows similar steps as in Algorithm 1. However, different from Algorithm 1, at the tth iteration, the agent knows the history $\mathcal{H}_{t-1,c,\hat{c}}$ of true contexts in addition to that of noisy contexts. Consequently, in the *de-noising* step, the agent evaluates the predictive posterior distribution as

$$P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathbb{E}_{P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})}[P(c_t|\hat{c}_t, \gamma^*)], \tag{28}$$

where $P(c_t|\hat{c}_t, \gamma^*)$ is as defined in (2) and posterior distribution $P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})$ is obtained via Baye's rule as $P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}}) \propto P(\gamma^*) \prod_{\tau=1}^{t-1} P(c_\tau, \hat{c}_\tau|\gamma^*)$ using the history of true and noisy contexts.

For the Gaussian context, noise as considered in Section 3.1, the predictive posterior distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathcal{N}(\tilde{V}_t, \tilde{R}_t^{-1})$ is multivariate Gaussian with the inverse covariance matrix,

$$\tilde{R}_t = M - \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1}, \tag{29}$$

and the mean vector

$$\tilde{V}_t = \tilde{R}_t^{-1}\Big(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}\hat{c}_t + \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1}\sum_{\tau=1}^{t-1}(\hat{c}_\tau - c_\tau) - \Sigma_n^{-1}\tilde{H}_t^{-1}\Sigma_n^{-1}M^{-1}(\Sigma_c^{-1}\mu_c - \Sigma_n^{-1}\hat{c}_t)\Big), \tag{30}$$

where $M = \Sigma_c^{-1} + \Sigma_n^{-1}$ and $\tilde{H}_t = \Sigma_n^{-1}M^{-1}\Sigma_n^{-1} + (t-1)\Sigma_n^{-1} + \Sigma_\gamma^{-1}$. Derivation can be found in Appendix B.2.4.

Following the denoising step, the next step in Algorithm 2 is a conventional Thompson sampling step, thanks to access to delayed true contexts. Consequently, the agent can evaluate the posterior distribution $P(\theta^*|\mathcal{H}_{t-1,r,a,c})$ with known contexts as in (9) and use it to sample $\theta_t \sim P(\theta^*|\mathcal{H}_{t-1,r,a,c})$. For 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 distribution whose inverse covariance matrix and mean, respectively, evaluate as

$$\tilde{\Sigma}_{t-1} = \frac{1}{\lambda}\mathbb{I} + \frac{1}{\sigma^2}\sum_{\tau=1}^{t-1}\phi(a_\tau, c_\tau)\phi(a_\tau, c_\tau)^\top \tag{31}$$

$$\tilde{\mu}_{t-1} = \frac{\tilde{\Sigma}_{t-1}^{-1}}{\sigma^2}\Big(\sum_{\tau=1}^{t-1} r_\tau\phi(a_\tau, c_\tau)\Big). \tag{32}$$

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

$$a_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta_t, \tag{33}$$

where we use the expected feature map $\psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}}) := \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{t-1,c,\hat{c}})}[\phi(a_t, c_t)]$.

Information-Theoretic Bayesian Regret Bounds

In this section, we derive an information-theoretic upper bound on the Bayesian regret (7) of Algorithm 2 for Gaussian CBs. The following theorem presents our main result.

Theorem 2. *Under Assumption 1 and assuming that covariance matrices satisfy* $\Sigma_\gamma \Sigma_n^{-1} \succ 0$, *the following inequality holds for* $\delta \in (0, 1)$ *when* $\lambda \leq \sigma^2$,

$$\mathcal{R}^T(\pi^{\mathrm{TS}}_{\mathrm{delay}}) \leq U(m, \lambda) + 4T\delta^2\sqrt{\frac{2m\lambda}{\pi}} + 2\sqrt{2\lambda mTd \log\Big(1 + T\mathrm{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big) \log\Big(\frac{2m}{\delta}\Big)},$$

where $U(m, \lambda)$ *is as defined in* (21).

Theorem 2 shows that Algorithm 2 achieves $\tilde{O}(m\sqrt{T})$ regret with the choice of $\delta = 1/T$ if $d = O(m)$. Furthermore, due to the absence of posterior mistmatch, the upper bound above is tighter than that of Theorem 1.

We now outline the main lemmas required to prove Theorem 2. To this end, we re-use the notation

$$\hat{a}_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^* \tag{34}$$

to define the optimal action maximizing the mean reward $\psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^*$.

To derive the regret upper bound in Theorem 2, we first decompose the Bayesian cumulative regret (7) of Algorithm 2 ($\pi^{\mathrm{TS}}_{\mathrm{delay}}$), similar to (23), into the following three terms,

$$\begin{aligned}
\mathcal{R}^T(\pi^{\mathrm{TS}}_{\mathrm{delay}}) &= \mathcal{R}^T_{\mathrm{d,CB}} + \mathcal{R}^T_{\mathrm{d,EE1}} + \mathcal{R}^T_{\mathrm{d,EE2}} \text{where,} \\
\mathcal{R}^T_{\mathrm{d,CB}} &= \sum_{t=1}^T \mathbb{E}\Big[\psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^* - \psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^*\Big], \\
\mathcal{R}^T_{\mathrm{d,EE1}} &= \sum_{t=1}^T \mathbb{E}\Big[\psi(a^*_t, \hat{c}_t|\gamma^*)^\top \theta^* - \psi(\hat{a}_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^*\Big], \\
\mathcal{R}^T_{\mathrm{d,EE2}} &= \sum_{t=1}^T \mathbb{E}\Big[\psi(a_t, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^* - \psi(a_t, \hat{c}_t|\gamma^*)^\top \theta^*\Big].
\end{aligned} \tag{35}$$

An upper bound on $\mathcal{R}^T(\pi^{\mathrm{TS}}_{\mathrm{delay}})$ can be then obtained by separately bounding each of the three terms in (35).

In (35), the first term $\mathcal{R}^T_{\mathrm{d,CB}}$ corresponds to the Bayesian cumulative regret of a standard contextual TS algorithm that uses $\psi(a, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^*$ for $a \in \mathcal{A}$ as the mean reward function. Note, that due to availability of delayed true contexts, there is no posterior mismatch in Algorithm 2. Hence, we apply Cor. 3 in [19] to yield the following upper bound on $\mathcal{R}^T_{\mathrm{d,CB}}$.

Lemma 3. *Under Assumption 1, the following upper bound on* $\mathcal{R}^T_{\mathrm{d,CB}}$ *holds for* $\frac{\lambda}{\sigma^2} \leq 1$,

$$\mathcal{R}^T_{\mathrm{d,CB}} \leq U(m, \lambda), \tag{36}$$

where $U(m, \lambda)$ *is defined as in* (21).

Lemma 3 gives a tighter bound in comparison to Lemma 1 where the posterior mismatch results in additional error terms in the regret bound.

We now upper bound the second term $\mathcal{R}^T_{\text{d,EE1}}$ of (35), which similar to the term $\mathcal{R}^T_{\text{EE1}}$ in (23), captures the error in approximating the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$ via the inferred predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})$. The following lemma shows that this 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. This bound also holds for the third term $\mathcal{R}^T_{\text{d,EE2}}$ of (35) which similarly accounts for the average approximation error.

Lemma 4. *Under Assumption 1, for any $\delta \in (0,1)$, we have the following upper bound,*

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

Furthermore, if the covariance matrices satisfy that $\Sigma_\gamma \Sigma_n^{-1} \succ 0$, we obtain that

$$I(\gamma^*; \mathcal{H}_{T,c,\hat{c}}) \leq \frac{d}{2}\log\Big(1 + T\text{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big)\Big).$$

Combining Lemmas 3 and 4 then gives us the upper bound on $\mathcal{R}^T(\pi^{\text{TS}}_{\text{delay}})$ in Theorem 1.

5. Experiments and Final Remarks

In this section, we experimentally validate the performance of the proposed algorithms on synthetic and real-world datasets. Details of implementation can be found in Appendux D.

Synthetic Datasets: For synthetic datasets, we go beyond Gaussian bandits and evaluate our algorithms for logistic contextual bandits (see Figure 1 (Left) and (Center)). In both these settings, we consider Gaussian contexts and context noise as in Section 3.1 with parameters $\Sigma_c = \mathbb{I}$, $\Sigma_\gamma = \sigma_\gamma^2\mathbb{I}$, $\Sigma_n = \sigma_n^2\mathbb{I}$ for some $\sigma_\gamma^2, \sigma_n^2 > 0$. We further consider action $a \in \mathcal{A}$ and context $c \in \mathcal{C}$ to be $d = 5$ dimensional vectors with a_i and c_i, respectively, denoting their ith component. We use $\phi(a,c) = [a_1^2, a_2^2, a_3^2, a_4^2, a_5^2, c_1^2, c_2^2, c_3^2, c_4^2, c_5^2, a_1c_1, a_2c_2, a_3c_3, a_4c_4, a_5c_5]$ as the $m = 15$ dimensional feature vector.

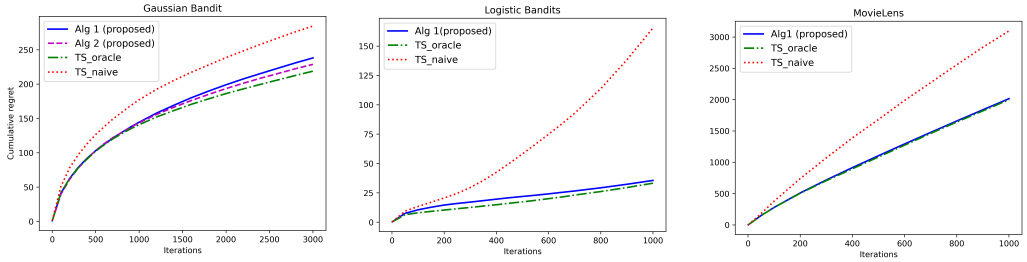

Figure 1. Comparison of Bayesian regret of proposed algorithms with baselines as a function of number of iterations. (**Left**): Gaussian bandits with $K = 40$, $\sigma_n^2 = \sigma_\gamma^2 = 1.1$; (**Center**) Logistic bandits with $K = 40$, $\sigma_n^2 = 2$, $\sigma_\gamma^2 = 2.5$; (**Right**) MovieLens dataset with added Gaussian context noise and Gaussian prior: parameters set as $\sigma_n^2 = 0.1$, $\sigma_\gamma^2 = 0.6$.

Gaussian Bandits: The mean reward function is given by $f(\theta^*, a, c) = \phi(a,c)^\top\theta^*$ with the feature map described above. Other parameters are fixed as $\sigma_\gamma^2 = \sigma_n^2 = 1.1$, $\sigma^2 = 2$ and $\lambda = 0.01$. Plots are averaged over 100 independent trials.

Logistic Bandits: The reward $r_t \in \{0,1\}$ is Bernoulli with mean reward given by $\mu(\phi(a,c)^\top\theta^*)$, where $\mu(z) = 1/(1+\exp(-z))$ is the sigmoid function. We consider a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbb{I})$ over θ^*. In Algorithm 1, we choose the sampling distribution

$$\bar{P}_t(\theta^*) \propto P(\theta^*)\prod_{\tau=1}^{t-1}\text{Ber}(\mu(\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\hat{c}})^\top\theta^*)).$$

However, the posterior $\bar{P}_t(\theta^*)$ is analytically intractable since Bernoulli reward-Gaussian prior forms a non-conjugate distribution pair. Consequently, we use Langevin Monte Carlo (LMC) [20] to sample from $\bar{P}_t(\theta^*)$. We run LMC for $I = 50$ iterations with learning rate $\eta_t = 0.2/t$ and inverse temperature $\beta^{-1} = 0.001$. Plots are averaged over 10 independent trials.

MovieLens Dataset: We use the MovieLens-100K dataset [21] to evaluate the performances. To utilise this dataset, we first perform non-negative matrix factorization on the rating matrix $R = [r_{c,a}] \in \mathbb{R}^{943\times1682}$ with 3 latent factors to obtain $R = WH$, where $W \in \mathbb{R}^{943\times3}$ and $H \in \mathbb{R}^{3\times1682}$. Each row vector W_c corresponds to an user context, while each column vector H_a corresponds to movie (action) features. The mean and variance of the Gaussian context distribution is estimated from the row vectors of W. We then add Gaussian noise to context as in the synthetic settings with $\sigma_n^2 = 0.1$.

We apply K-means algorithm to the column vectors of H to group the actions into $K = 20$ clusters. We use $m_k \in \mathbb{R}^3$ to denote the centroid and v_k to denote the variance of the kth cluster. We then fix the mean and variance of the Gaussian prior over θ^* as $\mu_\theta = (m_1, \ldots, m_K)$ and $\Sigma_\theta = \text{diag}(v_1\mathbb{I}_3, \ldots, v_K\mathbb{I}_3)$, with $\mathbb{I}_3$ denoting the 3×3 identity matrix, respectively. The feature vector $\phi(a, c)$ is then fixed as a 60-dimensional vector with vector W_c at the index of the cluster k to which action a belongs and zeros everywhere else. We further add mean-zero Gaussian noise to the mean reward $\phi(a, c)^\top\theta^*$ with variance $\sigma^2 = 0.01$. The Bayesian oracle in this experiment has access to the exact context noise parameter γ^* sampled from the Gaussian prior with variance $\Sigma_\gamma = \sigma_\gamma^2\mathbb{I}$, as well as the true θ^* sampled from the Gaussian prior $P(\theta^*)$.

Baselines: We compare our algorithms 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 [18], and can thus perform exact denoising via the predictive posterior distribution $P(c_t|\hat{c}, \gamma^*)$. This algorithm sets the benchmark for the best achievable regret.

Figure 1 (Left) corroborates our theoretical findings for Gaussian bandits. In particular, our algorithms (Algorithms 1 and 2) demonstrate sub-linear regret and achieve robust performance comparable to the best achievable performance of TS_oracle. We remark that while our regret analysis of Gaussian bandits is motivated due to the tractability of posterior distributions and the concentration properties of Gaussians, our empirical results for logistic bandits in Figure 1 (Center) show a promising extension of our algorithms to non-conjugate distributions. Extension of Bayesian regret analysis to such general distributions is left for future work. Further, our experiments on MovieLens data in Figure 1 (Right) validate the effectiveness of our algorithm in comparison to the benchmarks. The plot shows that our approach outperforms TS_naive and achieves comparable regret as that of TS_oracle which is the best achievable regret.

6. Conclusions

We studied a stochastic CB problem where the agent observes noisy contexts through a noise channel with an unknown channel parameter. For Gaussian bandits and Gaussian context noise, we introduced a TS algorithm that achieves $\tilde{O}(m\sqrt{T})$ Bayesian regret. The setting of Gaussian bandits with Gaussian noise was chosen for easy tractability of posterior distributions used in the proposed TS algorithms. 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.

Finally, we conjecture that our proposed modified TS algorithm and the information-theoretic Bayesian regret analysis could be extended to noisy contexts in multi-task bandit settings. In this regard, a good starting point would be to leverage prior works that study multi-armed hierarchical bandits [22] and contextual hierarchical bandits [23] with linear-Gaussian reward models. However, the critical challenge is to evaluate the posterior mismatch which requires a case-by-case analysis.

Author Contributions: Conceptualization, S.T.J. and S.M.; Methodology, S.T.J.; Formal analysis, S.T.J. and S.M.; Writing—original draft, S.T.J. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Kaggle, https://www.kaggle.com/datasets/prajitdatta/movielens-100k-dataset, accessed on 1 July 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Preliminaries

Definition A1 (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:*

$$\mathbb{E}_{P(y)}[\exp(\lambda(y - \mathbb{E}_{P(y)}[y])] \leq \exp\Big(\frac{\lambda^2 s^2}{2}\Big). \tag{A1}$$

Lemma A1 (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,*

$$|\mathbb{E}_{P(x)}[g(x)] - \mathbb{E}_{Q(x)}[g(x)]| \leq \sqrt{2s^2 D_{\mathrm{KL}}(P(x)\|Q(x))}. \tag{A2}$$

Proof. The inequality (A2) follows by using the Donsker-Varadhan inequality (20) with $f(x) = \lambda g(x)$ for $\lambda \in \mathbb{R}$. This yields that

$$\begin{aligned} D_{\mathrm{KL}}(P(x)\|Q(x)) &\geq \mathbb{E}_{P(x)}[\lambda g(x)] - \log \mathbb{E}_{Q(x)}[\exp(\lambda g(x))] \\ &\geq \mathbb{E}_{P(x)}[\lambda g(x)] - \mathbb{E}_{Q(x)}[\lambda g(x)] - \lambda^2 \frac{s^2}{2} \end{aligned} \tag{A3}$$

where the last inequality follows from the assumption of sub-Gaussianity. Rearranging, we obtain

$$\mathbb{E}_{P(x)}[\lambda g(x)] - \mathbb{E}_{Q(x)}[\lambda g(x)] \leq \lambda^2 \frac{s^2}{2} + D_{\mathrm{KL}}(P(x)\|Q(x)). \tag{A4}$$

For $\lambda > 0$, we obtain that

$$\mathbb{E}_{P(x)}[g(x)] - \mathbb{E}_{Q(x)}[g(x)] \leq \lambda \frac{s^2}{2} + \frac{D_{\mathrm{KL}}(P(x)\|Q(x))}{\lambda}, \tag{A5}$$

and optimizing over $\lambda > 0$ then yields that

$$\mathbb{E}_{P(x)}[g(x)] - \mathbb{E}_{Q(x)}[g(x)] \leq \sqrt{2s^2 D_{\mathrm{KL}}(P(x)\|Q(x))}. \tag{A6}$$

Similarly, for $\lambda < 0$, we obtain that

$$\mathbb{E}_{Q(x)}[g(x)] - \mathbb{E}_{P(x)}[g(x)] \leq \sqrt{2s^2 D_{\mathrm{KL}}(P(x)\|Q(x))}. \tag{A7}$$

□

Lemma A2. *Let $x \in \mathbb{R}^n$ be distributed according to $Q(x) = \prod_{i=1}^n \mathcal{N}(x_i|\mu_i, \sigma_i^2)$, i.e., each element of the random vector is independently distributed according to a Gaussian distribution with mean μ_i and variance σ_i^2. Let $g(x) = \max_{i=1,\ldots n} x_i$ denote the maximum of n Gaussian random variables. Then, the following inequality holds for $\lambda \geq 0$,*

$$\log \mathbb{E}_{Q(x)}[\exp(\lambda g(x))] \leq \log n + \lambda \max_i \mu_i + \lambda^2 \frac{\max_i \sigma_i^2}{2}. \tag{A8}$$

For any distribution $P(x)$ that is absolutely continuous with respect to $Q(x)$, we then have the following change of measure inequality,

$$\mathbb{E}_{P(x)}[g(x)] - \mathbb{E}_{Q(x)}[g(x)] \leq \sqrt{2\Big(\log n + D_{\mathrm{KL}}(P(x)\|Q(x))\Big) \max_i \sigma_i^2}. \tag{A9}$$

Proof. The proof of inequality (A8) follows from standard analysis (see [14]). We present it here for the sake of completeness. The following sequence of relations hold for any $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{E}_{Q(x)}[\exp(\lambda g(x))] = \mathbb{E}_{Q(x)}[\max_i \exp(\lambda x_i)] &\leq \sum_{i=1}^{n} \mathbb{E}_{Q(x_i)}[\exp(\lambda x_i)] \\
&= \sum_{i=1}^{n} \exp\Big(\lambda \mu_i + \lambda^2 \sigma_i^2/2\Big) \\
&\leq n \exp(\lambda \max_i \mu_i + \lambda^2 \max_i \sigma_i^2/2\Big).
\end{aligned} \tag{A10}$$

Taking logarithm on both sides of the inequality yields the upper bound in (A8). We now apply the DV inequality (20) as in (A3). This yields that

$$\begin{aligned}
D_{\mathrm{KL}}(P(x)\|Q(x)) &\geq \mathbb{E}_{P(x)}[\lambda g(x)] - \log \mathbb{E}_{Q(x)}[\exp(\lambda g(x))] \\
&\overset{(a)}{\geq} \mathbb{E}_{P(x)}[\lambda g(x)] - \log n - \lambda \max_i \mu_i - \lambda^2 \frac{\max_i \sigma_i^2}{2} \\
&\overset{(b)}{\geq} \mathbb{E}_{P(x)}[\lambda g(x)] - \mathbb{E}_{Q(x)}[\lambda g(x)] - \log n - \lambda^2 \frac{\max_i \sigma_i^2}{2},
\end{aligned} \tag{A11}$$

where the inequality in (a) follows from (A8). The inequality in (b) follows from observing that $\max_i x_i \geq x_i$ for all i, whereby we obtain that $\mathbb{E}_{Q(x)}[\max_i x_i] \geq \mu_i$ which holds for all i. The latter inequality implies that $\mathbb{E}_{Q(x)}[\max_i x_i] \geq \max_i \mu_i$. Re-arranging and optimizing over $\lambda \geq 0$ then yields the required inequality in (A9). □

Appendix B. Linear-Gaussian Contextual Bandits with Delayed Contexts

In this section, we provide all the details relevant to the Bayesian cumulative regret analysis of TS for delayed, linear-Gaussian contextual bandits.

Appendix B.1. TS Algorithm for Linear-Gaussian Bandits with Delayed True Contexts

The pseudocode for the TS algorithm for Gaussian bandits is given in Algorithm A1.

Algorithm A1: TS with Delayed Contexts for Gaussian Bandits ($\pi^{\mathrm{TS}}_{\mathrm{delay}}$)

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 (29) and (30) 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 (32) and (31).
7: Agent chooses action a_t as in (33) using θ_t and $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$.
8: Agent observes reward r_t corresponding to a_t, and the true context c_t.
9: **end for**

Appendix B.2. Derivation of Posterior and Predictive Posterior Distributions

In this section, we provide detailed derivation of posterior predictive distribution for Gaussian bandits. To this end, we first derive the exact predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$.

Appendix B.2.1. Derivation of $P(c_t|\hat{c}_t, \gamma^*)$

We begin by noting that

$$\begin{aligned} P(c_t|\hat{c}_t, \gamma^*) = \frac{P(c_t)P(\hat{c}_t|c_t, \gamma^*)}{P(\hat{c}_t|\gamma^*)} &\propto P(c_t)P(\hat{c}_t|c_t, \gamma^*) \\ &= \mathcal{N}(\mu_c, \Sigma_c)\mathcal{N}(c_t + \gamma^*, \Sigma_n). \end{aligned}$$

Subsequently,

$$\begin{aligned} \log(P(c_t|\hat{c}_t, \gamma^*)) &\propto \log(P(c_t)P(\hat{c}_t|c_t, \gamma^*)) \\ &\propto -\frac{1}{2}\Big((c_t - \mu_c)^\top \Sigma_c^{-1}(c_t - \mu) + (\hat{c}_t - c_t - \gamma^*)^\top \Sigma_n^{-1}(\hat{c}_t - c_t - \gamma^*)\Big) \\ &= -\frac{1}{2}\Big(c_t^\top(\Sigma_c^{-1} + \Sigma_n^{-1})c_t - \Big(\mu_c^\top \Sigma_c^{-1} + (\hat{c}_t - \gamma^*)^\top \Sigma_n^{-1}\Big)c_t \\ &\quad - c_t^\top\Big(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}(\hat{c}_t - \gamma^*)\Big) + (\hat{c}_t - \gamma^*)^\top \Sigma_n^{-1}(\hat{c}_t - \gamma^*)\Big) \\ &= -\frac{1}{2}\Big(c_t^\top M c_t - A_t^\top M c_t - c_t^\top MA + A_t^\top MA - A_t^\top MA \\ &\quad + (\hat{c}_t - \gamma^*)^\top \Sigma_n^{-1}(\hat{c}_t - \gamma^*)\Big), \end{aligned} \tag{A12}$$

where we have defined

$$M = \Sigma_c^{-1} + \Sigma_n^{-1} \tag{A13}$$

$$A_t = (M^{-1})^\top\Big(\Sigma_c^{-1}\mu_c + \Sigma_n^{-1}(\hat{c}_t - \gamma^*)\Big). \tag{A14}$$

From (A12), we obtain

$$\log(P(c_t|\hat{c}_t, \gamma^*)) \propto -\frac{1}{2}\Big(c_t^\top M c_t - A_t^\top M c_t - c_t^\top MA + A_t^\top MA\Big).$$

This implies that

$$P(c_t|\hat{c}_t, \gamma^*) = \mathcal{N}(A_t, M^{-1}). \tag{A15}$$

Appendix B.2.2. Derivation of $P(\hat{c}_t|\gamma^*)$

We now derive the distribution $P(\hat{c}_t|\gamma^*)$ which is defined in (3) as

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

Hence, $P(\hat{c}_t|\gamma^*)$ can be obtained by marginalizing the joint distribution $P(c_t)P(\hat{c}_t|c_t, \gamma^*) = P(c_t|\hat{c}_t, \gamma^*)P(\hat{c}_t|\gamma^*)$ over c_t. To this end, we use (A12) to obtain,

$$\begin{aligned} \log(P(c_t)P(\hat{c}_t|c_t, \gamma^*)) &= \log(P(c_t|\hat{c}_t, \gamma^*)P(\hat{c}_t|\gamma^*)) \\ &\propto \log(P(c_t|\hat{c}_t, \gamma^*)) - \frac{1}{2}\Big(-A_t^\top MA + (\hat{c}_t - \gamma^*)^\top \Sigma_n^{-1}(\hat{c}_t - \gamma^*)\Big), \end{aligned}$$

which implies that

$$\begin{aligned}\log(P(\hat{c}_t|\gamma^*)) &\propto -\frac{1}{2}\Big(-A_t^\top MA + (\hat{c}_t-\gamma^*)^\top\Sigma_n^{-1}(\hat{c}_t-\gamma^*)\Big)\\ &\propto -\frac{1}{2}\Big(\hat{c}_t^\top\Big(\Sigma_n^{-1}-\Sigma_n^{-1}(M^{-1})^\top\Sigma_n^{-1}\Big)\hat{c}_t - \hat{c}_t^\top\Big(\Sigma_n^{-1}(M^{-1})^\top(-\Sigma_n^{-1}\gamma^*+\Sigma_c^{-1}\mu_c)\\ &\quad+\Sigma_n^{-1}\gamma^*\Big) - \Big((\mu_c^\top\Sigma_c^{-1}-\gamma^{*\top}\Sigma_n^{-1})(M^{-1})\Sigma_n^{-1}+\gamma^{*\top}\Sigma_n^{-1}\Big)\hat{c}\Big)\\ &= -\frac{1}{2}\Big(\hat{c}_t^\top G\hat{c}_t - F^\top G\hat{c}_t - G^\top F\hat{c}_t\Big)\\ &\propto -\frac{1}{2}\Big((\hat{c}_t-F)^\top G(\hat{c}_t-F)\Big),\end{aligned}$$

where

$$G = \Sigma_n^{-1}-\Sigma_n^{-1}(M^{-1})^\top\Sigma_n^{-1} \tag{A16}$$

$$F = (G^{-1})^\top\Big(\Sigma_n^{-1}(M^{-1})^\top(-\Sigma_n^{-1}\gamma^*+\Sigma_c^{-1}\mu_c)+\Sigma_n^{-1}\gamma^*\Big) = (G^{-1})^\top(G\gamma^*+\Sigma_n^{-1}(M^{-1})^\top\Sigma_c^{-1}\mu_c). \tag{A17}$$

Thus,

$$P(\hat{c}_t|\gamma^*) = \mathcal{N}(F, G^{-1}). \tag{A18}$$

Appendix B.2.3. Derivation of $P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})$

We now derive the posterior distribution $P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})$. To this end, we use Baye's theorem as

$$\begin{aligned}P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}}) &\propto \prod_{\tau=1}^{t-1}P(\hat{c}_\tau|c_\tau,\gamma^*)P(\gamma^*)\\ &= \prod_{\tau=1}^{t-1}\mathcal{N}(c_\tau+\gamma^*,\Sigma_n)\mathcal{N}(\mathbf{0},\Sigma_\gamma).\end{aligned}$$

We then have,

$$\begin{aligned}\log P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}}) &\propto -\frac{1}{2}\Big(\sum_{\tau=1}^{t-1}\Big((\hat{c}_\tau-c_\tau-\gamma^*)^\top\Sigma_n^{-1}(\hat{c}_\tau-c_\tau-\gamma^*)\Big)+\gamma^{*\top}\Sigma_\gamma^{-1}\gamma^*\Big)\\ &= -\frac{1}{2}\Big(\sum_{\tau=1}^{t-1}\Big((-\hat{c}_\tau+c_\tau+\gamma^*)^\top\Sigma_n^{-1}(-\hat{c}_\tau+c_\tau+\gamma^*)\Big)+\gamma^{*\top}\Sigma_\gamma^{-1}\gamma^*\Big)\\ &\propto -\frac{1}{2}\Big(\gamma^{*\top}((t-1)\Sigma_n^{-1}+\Sigma_\gamma^{-1})\gamma^* - \gamma^{*\top}\Sigma_n^{-1}(\sum_{\tau=1}^{t-1}(\hat{c}_\tau-c_\tau))\\ &\quad-(\sum_{\tau=1}^{t-1}(\hat{c}_\tau-c_\tau))^\top\Sigma_n^{-1}\gamma^*\Big).\end{aligned}$$

Consequently, we obtain that,

$$P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}}) = \mathcal{N}(\gamma^*|Y_t, W_t^{-1}) \tag{A19}$$

where

$$W_t = (t-1)\Sigma_n^{-1} + \Sigma_\gamma^{-1} \tag{A20}$$

$$Y_t = (W_t^{-1})^\top \Sigma_n^{-1} \sum_{\tau=1}^{t-1} (\hat{c}_\tau - c_\tau). \tag{A21}$$

Appendix B.2.4. Derivation of Posterior Predictive Distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$

Using results from previous subsections, we are now ready to derive the posterior predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})$. Note, that $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathbb{E}_{P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})}[P(c_t|\hat{c}_t, \gamma^*)]$. We then have the following set of relations:

$$\begin{aligned}
&\log\Big(P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})P(c_t|\hat{c}_t,\gamma^*)\Big)\\
&\propto -\frac{1}{2}\Big((c_t - A_t)^\top M(c_t - A_t) + (\gamma^* - Y_t)^\top W_t(\gamma^* - Y_t)\Big)\\
&= -\frac{1}{2}\Big((c_t - D - E_t + (M^{-1})^\top\Sigma_n^{-1}\gamma^*)^\top M(c_t - D - E_t + (M^{-1})^\top\Sigma_n^{-1}\gamma^*)\\
&+ (\gamma^* - Y_t)^\top W_t(\gamma^* - Y_t)\Big)\\
&\propto -\frac{1}{2}\Big(\gamma^{*\top}(\Sigma_n^{-1}(M^{-1})^\top\Sigma_n^{-1} + W_t)\gamma^* - \gamma^{*\top}(\Sigma_n^{-1}(D + E_t - c_t) + W_tY_t) - ((D + E_t - c_t)^\top\Sigma_n^{-1}\\
&+ Y_t^\top W_t)\gamma^* + (c_t - D - E_t)^\top M(c_t - D - E_t)\Big)\\
&= -\frac{1}{2}\Big(\gamma^{*\top}\tilde{H}_t\gamma^* - \gamma^{*\top}\tilde{H}_t^\top\tilde{J}_t - \tilde{J}_t^\top\tilde{H}_t\gamma^* + \tilde{J}_t^\top\tilde{H}_t\tilde{J}_t - \tilde{J}_t^\top\tilde{H}_t\tilde{J}_t + (c_t - D - E_t)^\top M(c_t - D - E_t)\Big)\\
&\propto \log(P(\gamma^*|\mathcal{H}_{t,c,\hat{c}})) - \frac{1}{2}\Big(-\tilde{J}_t^\top\tilde{H}_t\tilde{J}_t + (c_t - D - E_t)^\top M(c_t - D - E_t)\Big)
\end{aligned}$$

where $D = (M^{-1})^\top\Sigma_c^{-1}\mu_c$, $E_t = (M^{-1})^\top\Sigma_n^{-1}\hat{c}_t$, $\tilde{H}_t = \Sigma_n^{-1}(M^{-1})^\top\Sigma_n^{-1} + W_t$ and $\tilde{J}_t = (\tilde{H}_t^{-1})^\top(\Sigma_n^{-1}(D + E_t - c_t) + W_tY_t)$.

Since $\log\Big(P(\gamma^*|\mathcal{H}_{t-1,c,\hat{c}})P(c_t|\hat{c}_t,\gamma^*)\Big) = \log\Big(P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})P(\gamma^*|\mathcal{H}_{t,c,\hat{c}})\Big)$, we obtain

$$\begin{aligned}
\log(P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}})) &\propto -\frac{1}{2}\Big(-\tilde{J}_t^\top\tilde{H}_t\tilde{J}_t + (c_t - D - E_t)^\top M(c_t - D - E_t)\Big)\\
&\propto -\frac{1}{2}\Big(c_t^\top\Big(M - \Sigma_n^{-1}(\tilde{H}_t^{-1})^\top\Sigma_n^{-1}\Big)c_t - c_t^\top\Big(M(D + E_t)\\
&- \Sigma_n^{-1}(\tilde{H}_t^{-1})^\top(\Sigma_n^{-1}(D + E_t) + W_tY_t)\Big) - \Big(M(D + E_t)\\
&- \Sigma_n^{-1}(\tilde{H}_t^{-1})^\top(\Sigma_n^{-1}(D + E_t) + W_tY_t)\Big)^\top c_t\Big).
\end{aligned}$$

This gives

$$P(c_t|\hat{c}_t, \mathcal{H}_{t-1,c,\hat{c}}) = \mathcal{N}(\tilde{V}_t, \tilde{R}_t^{-1}) \quad \text{where} \tag{A22}$$

$$\tilde{R}_t = M - \Sigma_n^{-1}(\tilde{H}_t^{-1})^\top\Sigma_n^{-1} \tag{A23}$$

$$\tilde{V}_t = (\tilde{R}_t^{-1})^\top\Big(M(D + E_t) - \Sigma_n^{-1}(\tilde{H}_t^{-1})^\top(\Sigma_n^{-1}(D + E_t) + W_tY_t)\Big). \tag{A24}$$

Appendix B.3. Proof of Lemma 3

We now present the proof of Lemma 3. To this end, we first recall that $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$ where $\mathcal{F}_t = \mathcal{H}_{t-1,r,a,c,\hat{c}} \cup \hat{c}_t$, and we denote $P_t(\theta^*)$ as the posterior distribution

$P(\theta^*|\mathcal{H}_{t-1,r,a,c})$ of θ^* given the history of observed reward-action-context tuples. We can then equivalently write $\mathcal{R}^T_{\text{d,CB}}$ as

$$\begin{aligned}\mathcal{R}^T_{\text{d,CB}} &= \sum_{t=1}^{T} \mathbb{E}\Big[\mathbb{E}_t\Big[\psi(\hat{a}_t,\hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^* - \psi(a_t,\hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top\theta^*\Big]\Big] \\ &= \sum_{t=1}^{T} \mathbb{E}\Big[\mathbb{E}_t\Big[\underbrace{f(\theta^*,\hat{a}_t,c_t) - f(\theta^*,a_t,c_t)}_{:=\Delta_t}\Big]\Big], \end{aligned} \tag{A25}$$

where $\psi(a,\hat{c}_t|\mathcal{H}_{c,\hat{c}}) = \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{t-1,c,\hat{c}})}[\phi(a,c_t)]$ is as defined in (33) and we have used $f(\theta^*,a_t,c_t) = \phi(a_t,c_t)^\top\theta^*$ to denote the mean-reward function. To obtain an upper bound on $\mathcal{R}^T_{\text{d,CB}}$, we define the following *lifted information ratio* as in [19],

$$\Gamma_t = \frac{(\mathbb{E}_t[\Delta_t])^2}{\Lambda_t} \tag{A26}$$

where

$$\Lambda_t = \mathbb{E}_t\Big[\Big(f(\theta^*,a_t,c_t) - \bar{f}(a_t,c_t)\Big)^2\Big], \tag{A27}$$

with $\bar{f}(a_t,c_t) = \mathbb{E}_t[f(\theta^*,a_t,c_t)|a_t,c_t]$ denoting the expectation of mean reward with respect to the posterior distribution $P_t(\theta^*)$. Subsequently, we obtain the following upper bound

$$\mathcal{R}^T_{\text{d,CB}} \le \sum_{t=1}^{T} \mathbb{E}\Big[\sqrt{\Gamma_t\Lambda_t}\Big] \le \sqrt{\mathbb{E}\Big[\sum_{t=1}^{T}\Gamma_t\Big]\Big[\sum_{t=1}^{T}\mathbb{E}\Big[\Lambda_t\Big]\Big]}, \tag{A28}$$

where the last inequality follows by an application of Cauchy–Schwarz inequality. An upper bound on $\mathcal{R}^T_{\text{d,CB}}$ then follows by obtaining an upper bound on the lifted information ratio Γ_t as well as on Λ_t.

We first evaluate the term Λ_t. To this end, note that $\bar{f}(a_t,c_t) = \phi(a_t,c_t)^\top\tilde{\mu}_{t-1}$, with $\tilde{\mu}_{t-1}$ defined as in (32). Using this, we obtain

$$\begin{aligned}\Lambda_t &= \mathbb{E}_t\Big[(\phi(a_t,c_t)^\top(\theta^* - \tilde{\mu}_{t-1}))^2\Big] \\ &= \mathbb{E}_t\Big[\phi(a_t,c_t)^\top(\theta^* - \tilde{\mu}_{t-1})(\theta^* - \tilde{\mu}_{t-1})^\top\phi(a_t,c_t)\Big] &\text{(A29)} \\ &= \mathbb{E}_t\Big[\phi(a_t,c_t)^\top\tilde{\Sigma}^{-1}_{t-1}\phi(a_t,c_t)\Big] = \mathbb{E}_t\Big[\|\phi(a_t,c_t)\|^2_{\tilde{\Sigma}^{-1}_{t-1}}\Big] &\text{(A30)}\end{aligned}$$

where $\tilde{\Sigma}_{t-1}$ is as in (31), and the third equality follows since conditional on $\mathcal{F}_t$, (a_t,c_t) is independent of θ^*. Subsequently, we can apply the elliptical potential lemma Lemma 19.4 in [24] using the assumption that $\|\phi(\cdot,\cdot)\|_2 \le 1$ and that $\sigma^2/\lambda \ge 1$. This results in

$$\begin{aligned}\sum_{t=1}^{T}\|\phi(a_t,c_t)\|^2_{\tilde{\Sigma}^{-1}_{t-1}} &= \sigma^2\sum_{t=1}^{T}\|\phi(a_t,c_t)\|^2_{(\sigma^2\tilde{\Sigma}_{t-1})^{-1}} \\ &\le 2\sigma^2\log\frac{\det(\tilde{\Sigma}_T)}{\det(\sigma^2/\lambda\mathbb{I})} = 2\sigma^2\Big(m\log\Big(\sigma^2/\lambda + T/m\Big) - m\log(\sigma^2/\lambda)\Big) \\ &= 2m\sigma^2\log\Big(1 + \frac{T\lambda}{m\sigma^2}\Big). \end{aligned} \tag{A31}$$

To upper bound the lifted information ratio term Γ_t, we can use Lemma 7 in [19]. To demonstrate how to leverage results from [19], we start by showing that the inequality $\Gamma_t \le m$ holds. To this end, we note that the lifted information ratio can be equivalently written as

$$\Gamma_t = \frac{\left(\mathbb{E}_t\left[f(\theta_t, a_t, c_t) - \bar{f}(a_t, c_t)\right]\right)^2}{\Lambda_t} \tag{A32}$$

which follows since

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &= \mathbb{E}_t\left[f(\theta^*, \hat{a}_t, c_t) - f(\theta^*, a_t, c_t)\right] \\
&= \mathbb{E}_t\left[f(\theta^*, \hat{a}_t, c_t) - \bar{f}(a_t, c_t)\right] \\
&= \sum_{a'} P_t(\hat{a}_t = a')\mathbb{E}_t\left[f(\theta^*, a', c_t)|\hat{a}_t = a'\right] - \sum_{a'} P_t(a_t = a')\mathbb{E}_t\left[\bar{f}(a_t = a', c_t)\right] \\
&= \sum_{a'} P_t(\hat{a}_t = a')\Big(\mathbb{E}_t\left[f(\theta^*, a', c_t)|\hat{a}_t = a'\right] - \mathbb{E}_t[\bar{f}(a', c_t)]\Big),
\end{aligned} \tag{A33}$$

where the second equality holds since conditioned on $\mathcal{F}_t$, (a_t, c_t) and θ^* are independent. In the third equality, we denote $P_t(\hat{a}_t) = P(\hat{a}_t|\mathcal{F}_t)$ and $P_t(a_t) = P(a_t|\mathcal{F}_t)$. Using these, the last equality follows since $a_t \overset{d}{=} \hat{a}_t$, i.e, $P_t(a_t) = P_t(\hat{a}_t)$. Now, let us define a $K \times K$ matrix M with entries given by

$$M_{a,a'} = \sqrt{P_t(\hat{a}_t = a')P_t(a_t = a)}\Big(\mathbb{E}_t\left[f(\theta^*, a, c_t)|\hat{a}_t = a'\right] - \mathbb{E}_t[\bar{f}(a, c_t)]\Big). \tag{A34}$$

Using this and noting that $P_t(\hat{a}_t) = P_t(a_t)$, we obtain that $\mathbb{E}_t[\Delta_t] = \mathrm{Tr}(M)$. We now try to bound Λ_t in terms of the matrix M. To see this, we can equivalently write Λ_t as

$$\begin{aligned}
\Lambda_t &= \sum_a P_t(a_t = a)\mathbb{E}_t\left[\Big(f(\theta^*, a, c_t) - \bar{f}(a, c_t)\Big)^2\right] \\
&= \sum_{a,a'} P_t(a_t = a)P_t(\hat{a}_t = a')\mathbb{E}_t\left[\Big(f(\theta^*, a, c_t) - \bar{f}(a, c_t)\Big)^2|\hat{a}_t = a'\right] \\
&\geq \sum_{a,a'} P_t(a_t = a)P_t(\hat{a}_t = a')\Big(\mathbb{E}_t\left[f(\theta^*, a, c_t) - \bar{f}(a, c_t)|\hat{a}_t = a'\right]\Big)^2 \\
&= \sum_{a,a'} P_t(a_t = a)P_t(\hat{a}_t = a')\Big(\mathbb{E}_t[f(\theta^*, a, c_t)|\hat{a}_t = a'] - \mathbb{E}_t[\bar{f}(a, c_t)]\Big)^2 = \|M\|_F^2,
\end{aligned} \tag{A35}$$

where the inequality follows by the application of Jensen's inequality. We thus obtain that

$$\Gamma_t \leq \frac{\mathrm{Tr}(M)^2}{\|M\|_F^2} \leq m,$$

where the last inequality follows from Prop. 5 in [25]. From Lemma 3 in [19], we also obtain $\Gamma_t \leq 2\log(1+K)$. This results in the upper bound $\Gamma_t \leq \min\{m, 2\log(1+K)\}$.

Using this and the bound of (A31) in (A28), we obtain

$$\mathcal{R}^T_{\mathrm{d,CB}} \leq \sqrt{2Tm\sigma^2 \min\{m, 2(1+\log K)\}\log\Big(1 + \frac{T\lambda}{m\sigma^2}\Big)}. \tag{A36}$$

Appendix B.4. Proof of Lemma 4

We now prove an upper bound on the term $\mathcal{R}^T_{\mathrm{d,EE1}}$. To this end, let us define the following event:

$$\mathcal{E} := \left\{\|\theta^*\|_2 \leq \sqrt{2\lambda m \log\Big(\frac{2m}{\delta}\Big)} := U\right\}. \tag{A37}$$

Note, that since $\theta^* \sim \mathcal{N}(\theta^*|\mathbf{0}, \lambda\mathbb{I})$, we obtain that with probability at least $1-\delta$, the following inequality holds $\|\theta^*\|_\infty \leq \sqrt{2\lambda \log\left(\frac{2m}{\delta}\right)}$. Since $\|\theta^*\|_2 \leq \sqrt{m}\|\theta^*\|_\infty$, the above inequality in turn implies the event $\mathcal{E}$ such that $P(\mathcal{E}) \geq 1-\delta$.

$$\begin{aligned}
\mathcal{R}^T_{\text{d,EE1}} &\overset{(a)}{\leq} \sum_{t=1}^{T} \mathbb{E}\Big[\psi(a_t^*, \hat{c}_t|\gamma^*)^\top \theta^* - \psi(a_t^*, \hat{c}_t|\mathcal{H}_{c,\hat{c}})^\top \theta^*\Big] \\
&= \sum_{t=1}^{T} \mathbb{E}\Big[\underbrace{\mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t^*, c_t)^\top \theta^*] - \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})}[\phi(a_t^*, c_t)^\top \theta^*]}_{:=\Delta\left(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\right)}\Big], \\
&= \sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}] + \sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}^c\}] \qquad \text{(A38)} \\
&\overset{(b)}{\leq} \sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}] + 2\delta T\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c], \qquad \text{(A39)}
\end{aligned}$$

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

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

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:

$$\begin{aligned}
\mathbb{E}[\|\theta^*\|_2|\mathcal{E}^c] &\overset{(a)}{\leq} \sqrt{m}\mathbb{E}[\|\theta^*\|_\infty|\mathcal{E}^c] \overset{(b)}{=} \sqrt{m}\mathbb{E}[\|\theta^*\|_\infty|\|\theta^*\|_\infty > u] \\
&= \sqrt{m}\sum_{i=1}^{m} P(\|\theta^*\|_\infty = |\theta_i^*|)\mathbb{E}\Big[\|\theta^*\|_\infty\Big|\|\theta^*\|_\infty = |\theta_i^*|, \|\theta^*\|_\infty > u\Big] \\
&\leq \sqrt{m}\sum_{i=1}^{m} \mathbb{E}\Big[|\theta_i^*|\Big||\theta_i^*| > u\Big] \overset{(c)}{=} 2\sqrt{m}\sum_{i=1}^{m}\int_{x>u} xg(x)dx \\
&\overset{(d)}{=} -2\lambda\sqrt{m}\sum_{i=1}^{m}\int_{x>u} g'(x)dx = 2\lambda m^{3/2} g(u) = 2\lambda m^{3/2}\frac{1}{\sqrt{2\pi\lambda}}\exp(-u^2/2\lambda) \\
&= \delta\sqrt{\frac{m\lambda}{2\pi}}, \qquad \text{(A40)}
\end{aligned}$$

where (a) follows since $\|\theta\|_2 \leq \sqrt{m}\|\theta\|_\infty$, (b) follows since $\|\theta^*\|_2 > \sqrt{m}\sqrt{2\lambda\log\left(\frac{2m}{\delta}\right)}$ implies that $\|\theta^*\|_\infty > \sqrt{2\lambda\log\left(\frac{2m}{\delta}\right)} := u$. 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 $2g(\theta_i^*)$ where $g(\theta_i^*) = \frac{1}{\sqrt{2\pi\lambda}}\exp(-\theta_i^{*2}/(2\lambda))$ is 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. Thus, we have the following upper bound

$$\sum_{t=1}^{T} \mathbb{E}[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}^c\}] \leq 2T\delta^2\sqrt{\frac{m\lambda}{2\pi}}. \tag{A41}$$

We now obtain an upper bound on $\sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big]$. To this end, note that

$$\begin{aligned}\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big] &\leq P(\mathcal{E})\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)|\mathcal{E}\Big] \\ &\leq \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)|\mathcal{E}\Big]. \end{aligned} \tag{A42}$$

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 A1 gives the following upper bound

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*), P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}] &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\sqrt{2U^2 D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})}\Big] \\ &\leq \sqrt{2TU^2\sum_{t=1}^{T}\mathbb{E}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big]} \\ &\overset{(a)}{=} \sqrt{2TU^2\sum_{t=1}^{T} I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})} \\ &\overset{(b)}{\leq} \sqrt{2TU^2\sum_{t=1}^{T} I(c_t,\hat{c}_t;\gamma^*|\mathcal{H}_{c,\hat{c}})} \qquad \text{(A43)} \\ &\overset{(c)}{=} \sqrt{2TU^2 I(\mathcal{H}_{T,c,\hat{c}};\gamma^*)} \qquad \text{(A44)}\end{aligned}$$

where the equality in (a) follows by the definition of condition mutual information

$$I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}}) := \mathbb{E}\Big[D_{\mathrm{KL}}\Big(P(c_t,\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\|P(c_t|\hat{c}_t,\mathcal{H}_{c,\hat{c}})P(\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})\Big)\Big],$$

and inequality in (b) follows since $I(c_t,\hat{c}_t;\gamma^*|\mathcal{H}_{c,\hat{c}}) = I(\hat{c}_t;\gamma^*|\mathcal{H}_{c,\hat{c}}) + I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}}) \geq I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{c,\hat{c}})$ due to the non-negativity of mutual information, and finally, the equality in (c) follows from the chain rule of mutual information.

We now analyze the mutual information $I(\gamma^*;\mathcal{H}_{T,c,\hat{c}})$ which can be written as

$$\begin{aligned}I(\gamma^*;\mathcal{H}_{T,c,\hat{c}}) &= H(\gamma^*) - H(\gamma^*|\mathcal{H}_{T,c,\hat{c}}) \\ &= \frac{1}{2}\log\Big((2\pi e)^d\det(\Sigma_\gamma)\Big) - \frac{1}{2}\log\Big((2\pi e)^d\det(W^{-1})\Big) \qquad \text{(A45)} \\ &= \frac{1}{2}\log\frac{\det(\Sigma_\gamma)}{\det(W^{-1})} \qquad \text{(A46)}\end{aligned}$$

where $W = (T-1)\Sigma_n^{-1} + \Sigma_\gamma^{-1} = \Sigma_\gamma^{-1}(\mathbb{I} + (T-1)\Sigma_\gamma\Sigma_n^{-1})$. Using this, we can equivalently write

$$\begin{aligned}I(\gamma^*;\mathcal{H}_{T,c,\hat{c}}) &= \frac{1}{2}\log\frac{1}{\det((\mathbb{I}+(T-1)\Sigma_\gamma\Sigma_n^{-1})^{-1})} \qquad \text{(A47)} \\ &= \frac{1}{2}\log\det(\mathbb{I}+(T-1)\Sigma_\gamma\Sigma_n^{-1}). \qquad \text{(A48)}\end{aligned}$$

Under the assumption that $\Sigma_\gamma \Sigma_n^{-1} \succ 0$, we have $\mathbb{I} + (T-1)\Sigma_\gamma \Sigma_n^{-1} \succ 0$, whereby using the determinant-trace inequality we obtain,

$$\begin{aligned} I(\gamma^*; \mathcal{H}_{T,c,\hat{c}}) &= \frac{1}{2} \log\Big(\det(\mathbb{I} + (T-1)\Sigma_\gamma \Sigma_n^{-1})\Big) \\ &\leq \frac{d}{2} \log\Big(\mathrm{Tr}(\mathbb{I} + (T-1)\Sigma_\gamma \Sigma_n^{-1})/d)\Big) \\ &= \frac{d}{2} \log\Big(1 + (T-1)\mathrm{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big) \\ &\leq \frac{d}{2} \log\Big(1 + T\mathrm{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big). \end{aligned}$$

Using this in (A44), we obtain that

$$\sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t, \gamma^*), P(c_t|\hat{c}_t, \mathcal{H}_{c,\hat{c}})\Big)\mathbb{I}\{\mathcal{E}\}\Big] \leq \sqrt{Td \log\Big(1 + T\mathrm{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big) U^2}. \tag{A49}$$

Finally, using this in (A39), gives the following upper bound

$$\mathcal{R}^T_{\mathrm{d,EE1}} \leq \sqrt{2\lambda m T d \log\Big(1 + T\mathrm{Tr}(\Sigma_\gamma \Sigma_n^{-1})/d\Big) \log\Big(\frac{2m}{\delta}\Big)} + 2T\delta^2 \sqrt{\frac{m\lambda}{2\pi}}. \tag{A50}$$

We finally note that same upper bound holds for the term $\mathcal{R}^T_{\mathrm{d,EE2}}$.

Appendix C. Linear-Gaussian Noisy Contextual Bandits with Unobserved True Contexts

Appendix C.1. Derivation of Posterior Predictive Distribution

In this section, we derive the posterior predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$ for Gaussian bandits with Gaussian context noise. To this end, we first derive the posterior $P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})$.

Appendix C.1.1. Derivation of Posterior $P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})$

Using Baye's theorem, we have

$$P(\gamma^*|\mathcal{H}_{t-1,\hat{c}}) \propto P(\gamma^*) \prod_{\tau=1}^{t-1} P(\hat{c}_\tau|\gamma^*),$$

where $P(\hat{c}_\tau|\gamma^*)$ is derived in (A18). Subsequently, we have that

$$\begin{aligned} \log p(\gamma^*|\mathcal{H}_{t-1}, \hat{c}) &\propto -\frac{1}{2} \sum_{\tau=1}^{t-1} \Big((\hat{c}_\tau - F)^\top G(\hat{c}_\tau - F)\Big) - \frac{1}{2}\Big(\gamma^{*\top} \Sigma_\gamma^{-1} \gamma^*\Big) \\ &\propto -\frac{1}{2}\Big(\gamma^{*\top}((t-1)G + \Sigma_\gamma^{-1})\gamma^* - \gamma^{*\top}(G\hat{c}_{1:t-1} - (t-1)\Sigma_n^{-1}(M^{-1})^\top \Sigma_c^{-1}\mu_c) \\ &- \Big(\hat{c}_{1:t-1}^\top G - (t-1)\mu_c^\top \Sigma_c^{-1} M^{-1} \Sigma_n^{-1}\Big)\gamma^*\Big), \end{aligned}$$

where we have denoted $\sum_{\tau=1}^{t-1} \hat{c}_\tau = \hat{c}_{1:t-1}$. We then obtain

$$P(\gamma^*|\mathcal{H}_{t-1,\hat{c}}) = \mathcal{N}(\tilde{M}_t, N_t^{-1}) \quad \text{where,} \tag{A51}$$

$$N_t = (t-1)G + \Sigma_\gamma^{-1} \tag{A52}$$

$$\tilde{M}_t = (N_t^{-1})^\top \Big(G\hat{c}_{1:t-1} - (t-1)\Sigma_n^{-1}(M^{-1})^\top \Sigma_c^{-1}\mu_c\Big). \tag{A53}$$

Appendix C.1.2. Derivation of $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$

The derivation of posterior predictive distribution $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}})$ follows in a similar line as that in Appendix B.2.4. We start the derivation by noting that $P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}}) = \mathbb{E}_{P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})}[P(c_t|\hat{c}_t, \gamma^*)]$.

We have

$$\begin{aligned}
&\log(P(\gamma^*|\mathcal{H}_{t-1,\hat{c}})P(c_t|\hat{c}_t,\gamma^*)) \\
&\propto -\frac{1}{2}\Big((c_t - A_t)^\top M(c_t - A_t) + (\gamma^* - \tilde{M}_t)^\top N(\gamma^* - \tilde{M}_t)\Big) \\
&= -\frac{1}{2}\Big(\Big(M^{-1})^\top\Sigma_n^{-1}\gamma^* + c_t - D - E_t\Big)^\top M\Big((M^{-1})^\top\Sigma_n^{-1}\gamma^* + c_t - D - E_t\Big) \\
&+ (\gamma^* - \tilde{M}_t)^\top N_t(\gamma^* - \tilde{M}_t)\Big) \\
&= -\frac{1}{2}\Big((\gamma^* - J_t)^\top H_t(\gamma^* - J_t) - J_t^\top H_t J_t + (c_t - D - E_t)^\top M(c_t - D - E_t) + \tilde{M}_t^\top N_t\tilde{M}_t\Big), \qquad \text{(A54)}
\end{aligned}$$

where A_t is defined in (A14), M is defined in (A13), $\tilde{M}_t$ in (A53), N in (A52), $D = (M^{-1})^\top\Sigma_c^{-1}\mu_c$, $E_t = (M^{-1})^\top\Sigma_n^{-1}\hat{c}_t$, $H_t = \Sigma_n^{-1}(M^{-1})^\top\Sigma_n^{-1} + N_t$ and $J_t = (H_t^{-1})^\top\Big(\Sigma_n^{-1}(-c_t + D + E_t) + N_t\tilde{M}_t\Big)$.

Subsequently, we have

$$\begin{aligned}
\log p(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}}) &\propto -\frac{1}{2}\Big(-J_t^\top H_t J_t + (c_t - D - E_t)^\top M(c_t - D - E_t)\Big) \qquad \text{(A55)} \\
&\propto -\frac{1}{2}\Big(c_t^\top\Big(M - \Sigma_n^{-1}(H_t^{-1})^\top\Sigma_n^{-1}\Big)c_t - c_t^\top\Big(M(D + E_t) - \Sigma_n^{-1}(H_t^{-1})^\top L_t^\top\Big) \\
&- \Big((D + E_t)^\top M - L_t(H_t^{-1})\Sigma_n^{-1}\Big), \qquad \text{(A56)}
\end{aligned}$$

where $L_t = (D + E_t)^\top\Sigma_n^{-1} + \tilde{M}_t^\top N_t$. Thus, we have,

$$\begin{aligned}
P(c_t|\hat{c}_t, \mathcal{H}_{t-1,\hat{c}}) &= \mathcal{N}(c_t|V_t, R_t^{-1}), \quad \text{where} \qquad \text{(A57)} \\
R_t &= M - \Sigma_n^{-1}(H_t^{-1})^\top\Sigma_n^{-1} \qquad \text{(A58)} \\
V_t &= (R_t^{-1})^\top\Big(M(D + E_t) - \Sigma_n^{-1}(H_t^{-1})^\top L_t^\top\Big). \qquad \text{(A59)}
\end{aligned}$$

Appendix C.1.3. Evaluating the KL Divergence between the True Posterior $P_t(\theta^*)$ and Sampling Distribution $\bar{P}_t(\theta^*)$

In this subsection, we analyze the true posterior distribution $P_t(\theta^*) := P(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$ and the approximate sampling distribution $\bar{P}_t(\theta^*) := \bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$, and derive the KL divergence $D_{\mathrm{KL}}(P_t(\theta^*)\|\bar{P}_t(\theta^*))$ between them. To see this, note that from Bayes's theorem, we have the following joint probability distribution

$$\begin{aligned}
P(\theta^*, \mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}) &= P(\theta^*)\underbrace{\mathbb{E}_{P(\gamma^*)}\Big[\prod_{\tau=1}^{t-1}\mathbb{E}_{P(c_\tau)}[P(\hat{c}_\tau|c_\tau,\gamma^*)P(r_\tau|a_\tau,c_\tau,\theta^*)]\Big]}_{:=P(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a},\theta^*)} \qquad \text{(A60)} \\
&= P(\theta^*)P(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a},\theta^*) \qquad \text{(A61)}
\end{aligned}$$

whereby we obtain that

$$P_t(\theta^*) \propto P(\theta^*)P(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a},\theta^*). \qquad \text{(A62)}$$

In particular, for general feature maps $\phi(a, c)$, the distribution $P(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*)$ cannot be exactly evaluated, even under Gaussian assumptions, resulting in the posterior $P_t(\theta^*)$ to be intractable, in general.

In contrast to this, our approximate TS-algorithm scheme assumes the following joint probability distribution,

$$\bar{P}(\theta^*, \mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}) = P(\theta^*) \underbrace{\mathbb{E}_{P(\gamma^*)}\Big[\prod_{\tau=1}^{t-1} \mathbb{E}_{P(c_\tau)}[P(\hat{c}_\tau|c_\tau, \gamma^*)\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)]\Big]}_{:=\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*)} \tag{A63}$$

$$= P(\theta^*)\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*), \tag{A64}$$

where

$$\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*) = \mathcal{N}(\psi(a_\tau, \hat{c}_\tau|\mathcal{H}_{\tau-1,\hat{c}}), \sigma^2).$$

Consequently, we have

$$\begin{aligned}\bar{P}_t(\theta^*) &\propto P(\theta^*)\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*) \\ &= P(\theta^*)\Big(\prod_{\tau=1}^{t-1} \bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)\Big)\Big(\mathbb{E}_{P(\gamma^*)}\Big[\prod_{\tau=1}^{t-1}[P(\hat{c}_\tau|\gamma^*)\Big]\Big).\end{aligned} \tag{A65}$$

As a result, we can upper bound the KL divergence $D_{\mathrm{KL}}(P_t(\theta^*)\|\bar{P}_t(\theta^*))$ as

$$\begin{aligned}
&D_{\mathrm{KL}}(P_t(\theta^*)\|\bar{P}_t(\theta^*)) \\
&\le D_{\mathrm{KL}}\Big(P(\theta^*, \mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a})\|\bar{P}(\theta^*, \mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a})\Big) \\
&= D_{\mathrm{KL}}\Big(P(\theta^*)P(\mathcal{H}_{t-1,r,\hat{c}}|\theta^*, \mathcal{H}_{t-1,a})\|P(\theta^*)\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*)\Big) \\
&= \mathbb{E}_{P(\theta^*)}\Big[D_{\mathrm{KL}}\Big(P(\mathcal{H}_{t-1,r,\hat{c}}|\theta^*, \mathcal{H}_{t-1,a})\|\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a}, \theta^*)\Big)\Big] \\
&\overset{(a)}{\le} \mathbb{E}_{P(\theta^*)P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[D_{\mathrm{KL}}\Big(\prod_{\tau=1}^{t-1} P(\hat{c}_\tau|c_\tau, \gamma^*)P(r_\tau|a_\tau, c_\tau, \theta^*)\|\prod_{\tau=1}^{t-1} P(\hat{c}_\tau|c_\tau, \gamma^*)\bar{P}(r_\tau|a_\tau, \mathcal{H}_{\tau,\hat{c}}, \theta^*)\Big)\Big] \\
&= \mathbb{E}_{P(\theta^*)}\mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[\sum_{\tau=1}^{t-1} \mathbb{E}_{P(\mathcal{H}_{\tau-1,\hat{c}}|\mathcal{H}_{\tau-1,c},\gamma^*)}\Big[D_{\mathrm{KL}}\Big(P(r_\tau|a_\tau, c_\tau, \theta^*)\|\bar{P}(r_\tau|a_\tau, \hat{c}_\tau, \mathcal{H}_{\tau-1,\hat{c}}, \theta^*)\Big)\Big]\Big] \\
&\overset{(b)}{=} \mathbb{E}_{P(\theta^*)}\mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[\sum_{\tau=1}^{t-1} \mathbb{E}_{P(\mathcal{H}_{\tau-1,\hat{c}}|\mathcal{H}_{\tau-1,c},\gamma^*)}\Big[\frac{(\phi(a_\tau, c_\tau)^\top\theta^* - \psi(\hat{c}_\tau, a_\tau|\mathcal{H}_{\tau-1,\hat{c}})^\top\theta^*)^2}{2\sigma^2}\Big]\Big], \\
&= \mathbb{E}_{P(\theta^*)}\mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[\sum_{\tau=1}^{t-1} \mathbb{E}_{P(\mathcal{H}_{\tau-1,\hat{c}}|\mathcal{H}_{\tau-1,c},\gamma^*)}\Big[\frac{|(\phi(a_\tau, c_\tau) - \psi(\hat{c}_\tau, a_\tau|\mathcal{H}_{\tau-1,\hat{c}}))^\top\theta^*|^2}{2\sigma^2}\Big]\Big], \\
&\overset{(c)}{\le} \mathbb{E}_{P(\theta^*)}\mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[\sum_{\tau=1}^{t-1} \mathbb{E}_{P(\mathcal{H}_{\tau-1,\hat{c}}|\mathcal{H}_{\tau-1,c},\gamma^*)}\Big[\frac{\|(\phi(a_\tau, c_\tau) - \psi(\hat{c}_\tau, a_\tau|\mathcal{H}_{\tau-1,\hat{c}})\|_2^2\|\theta^*\|_2^2}{2\sigma^2}\Big]\Big] \\
&\overset{(d)}{\le} \mathbb{E}_{P(\theta^*)}\mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}\Big[\sum_{\tau=1}^{t-1} \mathbb{E}_{P(\mathcal{H}_{\tau-1,\hat{c}}|\mathcal{H}_{\tau-1,c},\gamma^*)}\Big[\frac{4\|\theta^*\|_2^2}{2\sigma^2}\Big]\Big] \\
&= \sum_{\tau=1}^{t-1} \mathbb{E}_{P(\theta^*)}\Big[\frac{2\|\theta^*\|_2^2}{\sigma^2}\Big] \overset{(e)}{=} \frac{2(t-1)\lambda m}{\sigma^2}.
\end{aligned} \tag{A66}$$

In the above series of relationships,

- equality in (a) follows by noting that

$$\bar{P}(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a},\theta^*) = \mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}[\prod_{\tau=1}^{t-1} P(\hat{c}_\tau|c_\tau,\gamma^*)\bar{P}(r_\tau|a_\tau,\hat{c}_\tau,\theta^*)]$$

and

$$P(\mathcal{H}_{t-1,r,\hat{c}}|\mathcal{H}_{t-1,a},\theta^*) = \mathbb{E}_{P(\gamma^*)}\mathbb{E}_{P(\mathcal{H}_{t-1,c})}[\prod_{\tau=1}^{t-1} P(\hat{c}_\tau|c_\tau,\gamma^*)P(r_\tau|a_\tau,c_\tau,\theta^*)],$$

and applying Jensen's inequality on the jointly convex KL divergence,
- equality in (b) follows from evaluating the KL divergence between two Gaussian distributions with same variance σ^2 and with means $\phi(a_\tau,c_\tau)^\top\theta^*$ and $\psi(\hat{c}_\tau,a_\tau|\mathcal{H}_{\tau-1,\hat{c}})^\top\theta^*$ respectively,
- inequality in (c) follows from application of Cauchy–Schwarz inequality,
- inequality in (d) follows from Assumption 1,
- inequality in (e) follows from

$$\mathbb{E}_{P(\theta^*)}[\|\theta^*\|_2^2] = \sum_{j=1}^{m}\mathbb{E}[(\theta_j^*)^2] = \lambda m$$

since $P(\theta^*) = \mathcal{N}(\mathbf{0},\lambda\mathbb{I})$.

Appendix C.2. Proof of Lemma 1

For notational simplicity, throughout this section we use $\psi(a) := \psi(a,\hat{c}_t|\mathcal{H}_{\hat{c}}) = \mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{t-1,\hat{c}})}[\phi(a,c_t)]$ to denote the expected feature map. Furthermore, we use $\mathcal{F}_t = \mathcal{H}_{t-1,r,a,\hat{c}} \cup \hat{c}_t$.

We start by distinguishing the true and approximated posterior distributions. Recall that $P_t(\theta^*) := P(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$ denotes the true posterior and $\bar{P}_t(\theta^*) := \bar{P}(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}})$ denotes the approximated posterior. We then denote $P_t(\hat{a}_t,\theta^*) := P(\hat{a}_t,\theta^*|\mathcal{F}_t)$ as the distribution of $\hat{a}_t$ and θ^* conditioned on $\mathcal{F}_t$, while $\bar{P}_t(\hat{a}_t,\theta^*) := \bar{P}(\hat{a}_t,\theta^*|\mathcal{F}_t)$ denote the distribution of $\hat{a}_t$ and θ^* under the sampling distribution. Furthermore, we have that $\bar{P}_t(a_t,\theta^*) = \bar{P}_t(a_t)\bar{P}_t(\theta^*) = P_t(a_t)\bar{P}_t(\theta^*)$. We start by decomposing $\mathcal{R}_{\mathrm{CB}}^T$ into the following three differences,

$$\begin{aligned}
\mathcal{R}_{\mathrm{CB}}^T &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}_{P_t(\hat{a}_t,\theta^*)}[\psi(\hat{a}_t)^\top\theta^*] - \mathbb{E}_{P_t(a_t,\theta^*)}[\psi(a_t)^\top\theta^*]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\Big[\underbrace{\mathbb{E}_{\bar{P}_t(\hat{a}_t,\theta^*)}[\psi(\hat{a}_t)^\top\theta^*] - \mathbb{E}_{\bar{P}_t(a_t,\theta^*)}[\psi(a_t)^\top\theta^*]}_{:=\mathrm{Term}_1} + \underbrace{\mathbb{E}_{P_t(\hat{a}_t,\theta^*)}[\psi(\hat{a}_t)^\top\theta^*] - \mathbb{E}_{\bar{P}_t(\hat{a}_t,\theta^*)}[\psi(\hat{a}_t)^\top\theta^*]}_{:=\mathrm{Term}_2}] \\
&+ \underbrace{\mathbb{E}_{\bar{P}_t(a_t,\theta^*)}[\psi(a_t)^\top\theta^*] - \mathbb{E}_{P_t(a_t,\theta^*)}[\psi(a_t)^\top\theta^*]\Big]}_{:=\mathrm{Term}_3}
\end{aligned} \tag{A67}$$

We will separately upper bound each of the three terms in the above decomposition.

Appendix C.2.1. Upper Bound on Term_2

To obtain an upper bound on Term_2, note that the following equivalence holds $\mathbb{E}_{P_t(\hat{a}_t,\theta^*)}[\psi(\hat{a}_t)^\top\theta^*] = \mathbb{E}_{P_t(\theta^*)}\left[\max_a \psi(a)^\top\theta^*\right]$. Using this, we can rewrite Term_2 as

$$\mathrm{Term}_2 = \mathbb{E}_{P_t(\theta^*)}[\max_a \psi(a)^\top\theta^*] - \mathbb{E}_{\bar{P}_t(\theta^*)}[\max_a \psi(a)^\top\theta^*]. \tag{A68}$$

Note, here that when $\theta^* \sim \bar{P}_t(\theta^*)$, for each $a \in \mathcal{A}$, we have that $z_a = \psi(a)^\top \theta^*$ follows Gaussian distribution $\mathcal{N}(z_a | \psi(a)^\top \mu_{t-1}, \psi(a)^\top \Sigma_{t-1}^{-1} \psi(a))$ with mean $\psi(a)^\top \mu_{t-1}$ and variance $\psi(a)^\top \Sigma_{t-1}^{-1} \psi(a)$, where μ_{t-1} and Σ_{t-1} are as defined in (18) and (17), respectively. Thus, $\mathbb{E}_{\bar{P}_t(\theta^*)}[\max_a z_a]$ is the average of maximum of Gaussian random variables. We can then apply Lemma A2 with $P(x) = P_t(\theta^*)$, $Q(x) = \bar{P}_t(\theta^*)$, $n = |\mathcal{A}| = K$, $\mu_i = \psi(a)^\top \mu_{t-1}$ and $\sigma_i = \psi(a)^\top \Sigma_{t-1}^{-1} \psi(a)$ to obtain that

$$\text{Term}_2 \leq \sqrt{2(\log K + D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*)) \max_a \psi(a)^\top \Sigma_t^{-1} \psi(a)}. \tag{A69}$$

Using this, we obtain that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\text{Term}_2] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{2(\log K + D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*))) \max_a \psi(a)^\top \Sigma_t^{-1} \psi(a)}\right] \\
&\overset{(a)}{\leq} \sqrt{\left(\sum_{t=1}^{T} \mathbb{E}\left[2(\log K + D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*)))\right]\right)\left(\sum_{t=1}^{T} \mathbb{E}\left[\max_a \psi(a)^\top \Sigma_t^{-1} \psi(a)\right]\right)}, \\
&\overset{(b)}{\leq} \sqrt{2\lambda T \left(\sum_{t=1}^{T} \mathbb{E}\left[\log K + D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*))\right]\right)} \\
&\overset{(c)}{\leq} \sqrt{2\lambda T \left(T \log K + \sum_{t=1}^{T} 2(t-1)\frac{\lambda m}{\sigma^2}\right)} \\
&= \sqrt{2\lambda T^2 \log K + \frac{4\lambda^2 T(T^2 - T)m}{2\sigma^2}} \\
&\leq \sqrt{2\lambda T^2 \log K + \frac{2\lambda^2 T^3 m}{\sigma^2}},
\end{aligned} \tag{A70}$$

where the inequality in (a) follows from Cauchy–Schwarz inequality, and the inequality in (b) follows since

$$\sum_{t=1}^{T} \mathbb{E}\left[\max_a \psi(a)^\top \Sigma_t^{-1} \psi(a)\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\max_a \psi(a)^\top (\lambda \mathbb{I}) \psi(a)\right] \leq \lambda T,$$

which follows since $\Sigma_t^{-1} \leq \lambda \mathbb{I}$ and $\|\psi(a)\| \leq 1$. The inequality in (c) follows from (A66).

If $\lambda \leq \frac{\sigma^2}{T}$, we obtain that

$$\sum_{t=1}^{T} \mathbb{E}[\text{Term}_2] \leq \sqrt{2T\sigma^2(\log(K) + m)}. \tag{A71}$$

Appendix C.2.2. Upper Bound on Term_3

We can bound Term_3 by observing that

$$\text{Term}_3 = \mathbb{E}_{P_t(a_t)}[\psi(a_t)]^\top \left(\mathbb{E}_{\bar{P}_t(\theta^*)}[\theta^*] - \mathbb{E}_{P_t(\theta^*)}[\theta^*]\right) \tag{A72}$$

$$= \mathbb{E}_{\bar{P}_t(\theta^*)}[\Psi_t^\top \theta^*] - \mathbb{E}_{P_t(\theta^*)}[\Psi_t^\top \theta^*] \tag{A73}$$

where we used $\Psi_t = \mathbb{E}_{P_t(a_t)}[\psi(a_t)]$. Note, that for $\theta^* \sim \bar{P}_t(\theta^*)$, the random variable $\Psi_t^\top \theta^*$ is Gaussian with mean $\Psi_t^\top \mu_{t-1}$ and variance $\Psi_t^\top \Sigma_{t-1}^{-1} \Psi_t$. Consequently, $\Psi_t^\top \theta^*$ is also $\Psi_t^\top \Sigma_{t-1}^{-1} \Psi_t$-sub-Gaussian according to Definition A.1. By using Lemma A1, we then obtain that

$$|\text{Term}_3| \leq \sqrt{2(\Psi_t^\top \Sigma_t^{-1} \Psi_t) D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*))}. \tag{A74}$$

Using Cauchy–Schwarz inequality then yields that

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{T} |\text{Term}_3|] &\leq \sqrt{\Big(\sum_{t=1}^{T} \mathbb{E}[\Psi_t^\top \Sigma_t^{-1} \Psi_t]\Big)\Big(\sum_{t=1}^{T} \mathbb{E}[2D_{\text{KL}}(P_t(\theta^*) \| \bar{P}_t(\theta^*))]\Big)} \\ &\leq \sqrt{2\lambda T \frac{\lambda T^2 m}{\sigma^2}} = \sqrt{2\lambda^2 T^3 \frac{m}{\sigma^2}}.\end{aligned}$$

where the second inequality follows from (A66). As before, if $\lambda \leq \frac{\sigma^2}{T}$, we then obtain that

$$\mathbb{E}[\sum_{t=1}^{T} |\text{Term}_3|] \leq \sqrt{2Tm\sigma^2}. \tag{A75}$$

Appendix C.2.3. Upper Bound on Term_1

Note, that in Term_1, $\bar{P}_t(\hat{a}_t) = \bar{P}_t(a_t) = P_t(a_t)$, whereby the posterior is matched. Hence, one can apply bounds from conventional contextual Thompson Sampling here. For simplicity, we denote $\bar{\mathbb{E}}_t[\cdot] = \mathbb{E}_{\bar{P}}[\cdot|\mathcal{F}_t]$ to denote the expectation with respect to $\bar{P}_t(a, \theta)$. To this end, as in the proof of Lemma 3, we start by defining an information ratio,

$$\Gamma_t = \frac{{\text{Term}_1}^2}{\bar{E}_t\Big[\Big(\psi(a_t)^\top \theta^* - \psi(a_t)^\top \mu_t\Big)^2\Big]} := \Lambda_t, \tag{A76}$$

using which we obtain the upper bound on Term_1 as

$$\sum_{t=1}^{T} \mathbb{E}[\text{Term}_1] \leq \mathbb{E}\Big[\sum_{t=1}^{T} \sqrt{\Gamma_t \Lambda_t}\Big] \leq \sqrt{\Big(\sum_{t=1}^{T} \mathbb{E}[\Gamma_t]\Big)\Big(\sum_{t=1}^{T} \mathbb{E}[\Lambda_t]\Big)} \tag{A77}$$

by the Cauchy–Schwarz inequality.

Furthermore, we have

$$\begin{aligned}\Lambda_t = \bar{\mathbb{E}}_t\Big[(\psi(a_t)^\top(\theta^* - \mu_t))^2\Big] &= \bar{\mathbb{E}}_t\Big[\psi(a_t)^\top(\theta^* - \mu_{t-1})(\theta^* - \mu_{t-1})^\top \psi(a_t)\Big] \\ &= \bar{\mathbb{E}}_t[\psi(a_t)^\top \Sigma_{t-1}^{-1} \psi(a_t)] = \bar{E}_t[\|\psi(a_t)\|_{\Sigma_{t-1}^{-1}}],\end{aligned} \tag{A78}$$

where μ_{t-1} and Σ_{t-1} are defined as in (18) andd (17). Subsequently, using elliptical potential lemma, we obtain

$$\sum_{t=1}^{T} \Lambda_t \leq 2m\sigma^2 \log\Big(1 + \frac{(T)\lambda}{m\sigma^2}\Big). \tag{A79}$$

To obtain an upper bound on the information ratio Γ_t, we define $\bar{f}(\theta^*, a) = \psi(a)^\top \theta^*$ and $\bar{f}(a) = \psi(a)^\top \mu_{t-1}$ and let

$$M_{a,a'} = \sum_{a,a'} \sqrt{\bar{P}_t(a_t = a)\bar{P}_t(\hat{a}_t = a')}(\bar{\mathbb{E}}_t[\bar{f}(\theta^*, a)|\hat{a}_t = a'] - \bar{f}(a)). \tag{A80}$$

It is easy to see that

$$\begin{aligned}\text{Term}_1 &= \bar{\mathbb{E}}_t[\bar{f}(\hat{a}_t, \theta^*)] - \bar{\mathbb{E}}_t[\bar{f}(a_t)] \\ &= \sum_{a'} \bar{P}_t(\hat{a}_t = a') \Big(\bar{\mathbb{E}}_t[\bar{f}(\theta^*, a') | \hat{a}_t = a'] - \bar{\mathbb{E}}_t[\bar{f}(\hat{a}_t)] \Big) \\ &= \sum_{a'} \bar{P}_t(\hat{a}_t = a') \Big(\bar{\mathbb{E}}_t[\bar{f}(\theta^*, a') | \hat{a}_t = a'] - \bar{f}(a') \Big) = \text{Tr}(M), \end{aligned} \tag{A81}$$

where the second and last equality follows since $\bar{P}_t(a_t) = \bar{P}_t(\hat{a}_t)$. Similarly, we can relate Λ_t with the matrix $(M_{a,a'})$ as

$$\begin{aligned}\Lambda_t &= \bar{\mathbb{E}}_t[(\bar{f}(\theta^*, a_t) - \bar{f}(a_t))^2] \qquad \text{(A82)} \\ &= \sum_{a} \bar{P}_t(a_t = a) \bar{\mathbb{E}}_t[(\bar{f}(\theta^*, a) - \bar{f}(a))^2] \\ &= \sum_{a,a'} \bar{P}_t(a_t = a) \bar{P}_t(\hat{a}_t = a') \bar{\mathbb{E}}_t[(\bar{f}(\theta^*, a) - \bar{f}(a))^2 | \hat{a}_t = a'] \\ &\geq \sum_{a,a'} \bar{P}_t(a_t = a) \bar{P}_t(\hat{a}_t = a') \Big(\bar{\mathbb{E}}_t[(\bar{f}(\theta^*, a) - \bar{f}(a)) | \hat{a}_t = a'] \Big)^2 \\ &= \sum_{a,a'} \bar{P}_t(a_t = a) \bar{P}_t(\hat{a}_t = a') \Big(\bar{\mathbb{E}}_t[(\bar{f}(\theta^*, a) | \hat{a}_t = a'] - \bar{f}(a) \Big)^2 = \|M\|_F^2, \qquad \text{(A83)} \end{aligned}$$

whereby we obtain

$$\Gamma_t \leq \frac{\text{Tr}(M)^2}{\|M\|_F^2} \leq m, \tag{A84}$$

where the last inequality can be proved as in [25]. Following [19], it can be seen that $\Lambda_t \leq 2\log(1+K)$ also holds. Using this, together with the upper bound (A79) gives

$$\sum_{t=1}^{T} \mathbb{E}[\text{Term}_1] \leq \sqrt{2Tm\sigma^2 \min\{m, 2\log(1+K)\} \log\Big(1 + \frac{T\lambda}{m\sigma^2}\Big)}. \tag{A85}$$

Appendix C.3. Proof of Lemma 2

We first give an upper bound on the estimation error that does not require the assumption of a linear feature map.

Appendix C.3.1. A General Upper Bound on $\mathcal{R}_{\text{EE1}}^T$

To obtain an upper bound on $\mathcal{R}_{\text{EE1}}^T$ that does not require the assumption that $\phi(a, c) = G(a)c$, we leverage the same analysis as in the proof of Lemma 4. Subsequently, we obtain that

$$\begin{aligned}\mathcal{R}_{\text{EE1}}^T &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t | \hat{c}_t, \gamma^*), P(c_t | \hat{c}_t, \mathcal{H}_{\hat{c}}) \Big) \mathbf{1}\{\mathcal{E}\} \Big] + 2\delta T \mathbb{E}[\|\theta^*\|_2 | \mathcal{E}^c] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\Delta\Big(P(c_t | \hat{c}_t, \gamma^*), P(c_t | \hat{c}_t, \mathcal{H}_{\hat{c}}) \Big) \mathbf{1}\{\mathcal{E}\} \Big] + 2\delta^2 T \sqrt{\frac{m\lambda}{2\pi}}, \end{aligned}$$

where the event $\mathcal{E}$ is defined as in (A37). Subsequently, the first summation can be upper bounded as

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\Big[\Delta\Big(P(c_t|\hat{c}_t,\gamma^*),P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})\Big)\mathbf{1}\{\mathcal{E}\}\Big] &\le \sqrt{2TU^2\sum_{t=1}^{T}I(c_t;\gamma^*|\hat{c}_t,\mathcal{H}_{\hat{c}})}\\ &= \sqrt{2TU^2\sum_{t=1}^{T}(H(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})-H(c_t|\hat{c}_t,\gamma^*))}\\ &= \sqrt{TU^2\sum_{t=1}^{T}\log\Big(\det(R_t^{-1})\det(M_t)\Big)}\\ &\le \sqrt{TU^2(\mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM)^{-1})+\log(T)\mathrm{Tr}(\Sigma_c\Sigma_n^{-1}))} \qquad \text{(A86)}\end{aligned}$$

where U is defined as in (A37), $M_t=\Sigma_c^{-1}+\Sigma_n^{-1}$ and R_t is as in (13).

To derive the last inequality, we observe the following series of relationships starting from (13):

$$\begin{aligned}\Sigma_n^{-1}(H_t^{-1})^\top\Sigma_n^{-1}=(\Sigma_nH_t\Sigma_n)^{-1}&=\Big((t-1)\Sigma_n-(t-2)M^{-1}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_n\Big)^{-1}\\ R_t&=M-\Sigma_n^{-1}(H_t^{-1})^\top\Sigma_n^{-1}\\ &=M-\Big((t-1)\Sigma_n-(t-2)M^{-1}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_n\Big)^{-1}\\ &=M\Big[\mathbb{I}-M^{-1}\Big((t-1)\Sigma_n-(t-2)M^{-1}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_n\Big)^{-1}\Big]\\ R_t^{-1}&=\Big[\mathbb{I}-M^{-1}\Big((t-1)\Sigma_n-(t-2)M^{-1}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_n\Big)^{-1}\Big]^{-1}M^{-1}\\ &=\Big[\mathbb{I}-\Big((t-1)\Sigma_nM-(t-2)\mathbb{I}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM\Big)^{-1}\Big]^{-1}M^{-1}\end{aligned}$$

$$\begin{aligned}R_t^{-1}M&=\Big[\mathbb{I}-\Big((t-1)\Sigma_nM-(t-2)\mathbb{I}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM\Big)^{-1}\Big]^{-1}\\ &=\Big[\mathbb{I}-\Big((t-1)\Sigma_n\Sigma_c^{-1}+(t-1)\mathbb{I}-(t-2)\mathbb{I}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM\Big)^{-1}\Big]^{-1}\\ &=\Big[\mathbb{I}-\Big(\mathbb{I}+\underbrace{(t-1)\Sigma_n\Sigma_c^{-1}+\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM}_{:=P_t}\Big)^{-1}\Big]^{-1}\\ &\overset{(a)}{=}\Big[\mathbb{I}-\Big(\mathbb{I}-P_t(\mathbb{I}+P_t)^{-1}\Big)\Big]^{-1}=(P_t(\mathbb{I}+P_t)^{-1})^{-1}\end{aligned}$$

where the equality in (a) follows from Woodbury matrix identity and by the assumption that $\Sigma_n\Sigma_\gamma^{-1}\Sigma_nM\succ 0$, $\Sigma_n\Sigma_c^{-1}\succ 0$, we have $P_t\succ 0$ is invertible. Now,

$$\begin{aligned}\det(R_t^{-1}M)=\frac{1}{\det(P_t(\mathbb{I}+P_t)^{-1})}&=\det(P_t^{-1}(\mathbb{I}+P_t))\\ &=\det(P_t^{-1}+\mathbb{I})\\ &\overset{(b)}{\le}\left(\frac{\mathrm{Tr}(P_t^{-1}+\mathbb{I})}{d}\right)^d=(1+\mathrm{Tr}(P_t^{-1})/d)^d,\end{aligned}$$

where the inequality in (b) follows from the determinant-trace inequality. Subsequently, we have

$$\begin{aligned}\sum_{t=1}^{T}\log\Big(\det(R_t^{-1}M)\Big) &\leq \sum_{t=1}^{T} d\log(1+\mathrm{Tr}(P_t^{-1})/d)\\ &\overset{(c)}{\leq} \sum_{t=1}^{T}\mathrm{Tr}(P_t^{-1})\\ &= \mathrm{Tr}(P_1^{-1})+\sum_{t=2}^{T}\mathrm{Tr}(P_t^{-1})\\ &= \mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M)^{-1})+\sum_{t=2}^{T}\mathrm{Tr}(P_t^{-1})\\ &\overset{(d)}{\leq} \mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M)^{-1})+\log(T)\mathrm{Tr}(\Sigma_c\Sigma_n^{-1}),\end{aligned} \tag{A87}$$

where the inequality in (c) follows since $\log(1+x)\leq x$ for $x>0$ and the inequality in (d) follows since by assumption $P_t \succeq (t-1)\Sigma_n\Sigma_c^{-1}$, whereby we have $P_t^{-1}\preceq \frac{1}{t-1}\Sigma_c\Sigma_n^{-1}$ and consequently, $\mathrm{Tr}(P_t^{-1})\leq \frac{1}{t-1}\mathrm{Tr}(\Sigma_c\Sigma_n^{-1})$. Finally, we use that $\sum_{s=1}^{T}1/s\leq\log(T)$.

We thus obtain that

$$\mathcal{R}_{\mathrm{EE1}}^{T}\leq\sqrt{2\lambda mT\log(2m/\delta)\Big(\mathrm{Tr}((\Sigma_n\Sigma_\gamma^{-1}\Sigma_n M)^{-1})+\log(T)\mathrm{Tr}(\Sigma_c\Sigma_n^{-1})\Big)}+2\delta^2T\sqrt{\frac{m\lambda}{2\pi}} \tag{A88}$$

for $\delta\in(0,1)$. We note that same upper bound holds for the term $\mathcal{R}_{\mathrm{EE2}}^{T}$.

Appendix C.3.2. Upper Bound for Linear Feature Maps and Scaled Diagonal Covariance Matrices

We now obtain an upper bound on the estimation error under the assumption of a linear feature map $\phi(a,c)=G_a c$ such that $\|\phi(a,c)\|_2\leq 1$. The following set of inequalities hold:

$$\begin{aligned}\mathcal{R}_{\mathrm{EE1}}^{T} &= \sum_{t=1}^{T}\mathbb{E}\Big[\psi(a_t^*,\hat{c}_t|\gamma^*)^\top\theta^*-\psi(\hat{a}_t,\hat{c}_t|\mathcal{H}_{\hat{c}})^\top\theta^*\Big]\\ &\leq \sum_{t=1}^{T}\mathbb{E}\Big[\psi(a_t^*,\hat{c}_t|\gamma^*)^\top\theta^*-\psi(a_t^*,\hat{c}_t|\mathcal{H}_{\hat{c}})^\top\theta^*\Big]\\ &= \sum_{t=1}^{T}\mathbb{E}\Big[\mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t^*,c_t)^\top\theta^*]-\mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})}[\phi(a_t^*,c_t)^\top\theta^*].\end{aligned} \tag{A89}$$

Note, that $P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})=\mathcal{N}(c_t|V_t,R_t^{-1})$ where R_t and V_t are, respectively, defined in (13) and (14). Consequently, $\phi(a_t^*,c_t)^\top\theta^*=c_t^\top G_{a_t^*}^\top\theta^*$ is $s_t^2=\theta^{*\top}G_{a_t^*}R_t^{-1}G_{a_t^*}^\top\theta^*$-sub-Gaussian with respect to $P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})$. Consequently, using Lemma A1, we can upper bound the inner expectation of (A89) as

$$|\mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t^*,c_t)^\top\theta^*]-\mathbb{E}_{P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}})}[\phi(a_t^*,c_t)^\top\theta^*|\leq\sqrt{2s_t^2D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}}))}. \tag{A90}$$

Summing over t and using Cauchy–Schwarz inequality then gives

$$\mathcal{R}_{\mathrm{EE1}}^{T}\leq\sqrt{2(\sum_{t=1}^{T}\mathbb{E}[s_t^2])\Big(\sum_{t=1}^{T}\mathbb{E}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t,\gamma^*)\|P(c_t|\hat{c}_t,\mathcal{H}_{\hat{c}}))\Big]\Big)}. \tag{A91}$$

We now evaluate the KL-divergence term. To this end, note that conditioned on γ^* and $\hat{c}_t$, c_t is independent of $\mathcal{H}_{\hat{c}}$, i.e., $P(c_t|\hat{c}_t, \gamma^*, \mathcal{H}_{\hat{c}}) = P(c_t|\hat{c}_t, \gamma^*)$. This gives

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t, \gamma^*) \| P(c_t|\hat{c}_t, \mathcal{H}_{\hat{c}}))\Big] \\
&= \sum_{t=1}^{T} I(c_t; \gamma^*|\hat{c}_t, \mathcal{H}_{\hat{c}}) \\
&= \sum_{t=1}^{T} H(c_t|\hat{c}_t, \mathcal{H}_{\hat{c}}) - H(c_t|\hat{c}_t, \gamma^*) \\
&= \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{P(\hat{c}_t, \mathcal{H}_{\hat{c}})}[\log\det(R_t^{-1})] - \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{P(\hat{c}_t, \gamma^*)}[\log\det(M^{-1})] \\
&= \sum_{t=1}^{T} \frac{1}{2}\mathbb{E}_{P(\hat{c}_t, \mathcal{H}_{\hat{c}})}[\log\Big(\det(R_t^{-1})\det(M)\Big) \\
&\leq \frac{d\sigma_c^2}{\sigma_n^2}\Big(\frac{\sigma_\gamma^2}{\sigma_n^2 + \sigma_c^2} + \log(T-1)\Big), \qquad \text{(A92)}
\end{aligned}$$

where $M = \Sigma_c^{-1} + \Sigma_n^{-1}$ and R_t is as in (13). The first equality follows by noting that

$$\begin{aligned}
\mathbb{E}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t, \gamma^*) \| P(c_t|\hat{c}_t, \mathcal{H}_{\hat{c}}))\Big] &= \mathbb{E}_{P(\hat{c}_t, \mathcal{H}_{\hat{c}})}\Big[\mathbb{E}_{P(\gamma^*|\hat{c}_t, \mathcal{H}_{\hat{c}})}\Big[D_{\mathrm{KL}}(P(c_t|\hat{c}_t, \gamma^*, \mathcal{H}_{\hat{c}}) \| P(c_t|\hat{c}_t, \mathcal{H}_{\hat{c}}))\Big]\Big] \\
&= \mathbb{E}_{P(\hat{c}_t, \mathcal{H}_{\hat{c}})}[I(c_t; \gamma^*|\hat{c}_t, \mathcal{H}_{\hat{c}})]
\end{aligned}$$

with the outer expectation taken over $\hat{c}_t$ and $\mathcal{H}_{\hat{c}}$. The last inequality is proved in Appendix C.3.3 using that $\Sigma_c = \sigma_c^2\mathbb{I}$, $\Sigma_n = \sigma_n^2\mathbb{I}$ and $\Sigma_\gamma = \sigma_\gamma^2\mathbb{I}$.

We can now upper bound $\sum_t \mathbb{E}[s_t^2]$ as follows.

$$\begin{aligned}
\mathbb{E}[s_t^2] \leq \mathbb{E}[\max_a \theta^{*\top} G_a R_t^{-1} G_a^\top \theta^*] \leq \sum_a \mathbb{E}[\theta^{*\top} G_a R_t^{-1} G_a^\top \theta^*] &= \sum_a \mathrm{Tr}(G_a R_t^{-1} G_a^\top \mathbb{E}[\theta^* \theta^{*\top}]) \\
&= \lambda \sum_a \mathrm{Tr}(R_t^{-1} G_a^\top G_a) \\
&= \lambda b_t \mathrm{Tr}(\sum_a G_a^\top G_a)
\end{aligned}$$

where the last equality uses $R_t^{-1} = b_t\mathbb{I}$ as in (A95). Using (A95), we obtain

$$\begin{aligned}
\sum_t b_t &= \frac{\sigma_c^2\sigma_n^2}{f}\sum_t\Big(1 + \frac{\sigma_c^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2\sigma_n^2 + f\sigma_n^2}\Big) \\
&= \frac{\sigma_c^2\sigma_n^2 T}{f} + \sum_t \frac{\sigma_c^2}{f}\frac{\sigma_c^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2 + f} \\
&= \frac{\sigma_c^2\sigma_n^2 T}{f} + \frac{\sigma_c^4\sigma_\gamma^2}{f^2} + \sum_{t>1} \frac{\sigma_c^2}{f}\frac{\sigma_c^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2 + f} \\
&\leq \frac{\sigma_c^2\sigma_n^2}{f}T + \frac{\sigma_c^4\sigma_\gamma^2}{f^2} + \frac{\sigma_c^4}{f}\log(T-1).
\end{aligned}$$

Using the above relation, we obtain

$$\sum_t \mathbb{E}[s_t^2] = \lambda \mathrm{Tr}(\sum_a G_a^\top G_a)\sum_t b_t \leq \frac{\lambda K\sigma_c^2}{f}\max_a \mathrm{Tr}(G_a^\top G_a)\Big(\sigma_n^2 T + \frac{\sigma_c^2\sigma_\gamma^2}{f} + \sigma_c^2\log(T-1)\Big).$$

If $\lambda \leq \frac{d\sigma^2}{T}$, we obtain

$$\sum_t \mathbb{E}[s_t^2] \leq \frac{d\sigma^2 K\sigma_c^2}{f} \max_a \mathrm{Tr}(G_a^\top G_a)\Big(\sigma_n^2 + \frac{\sigma_c^2\sigma_\gamma^2}{Tf} + \sigma_c^2\frac{\log(T-1)}{T}\Big) = L$$

Using the above inequality together with (A92) in (A91) yields

$$\mathcal{R}_{\mathrm{EE1}}^T \leq \sqrt{2L\frac{d\sigma_c^2}{\sigma_n^2}\Big(\frac{\sigma_\gamma^2}{\sigma_n^2+\sigma_c^2} + \log(T-1)\Big)}. \tag{A93}$$

An upper bound on $\mathcal{R}_{\mathrm{EE2}}^T$ similarly follows.

Appendix C.3.3. Analysis of R_t for Scaled Diagonal Covariance Matrices

Assume that $\Sigma_n = \sigma_n^2\mathbb{I}$, $\Sigma_c = \sigma_c^2\mathbb{I}$ and $\Sigma_\gamma = \sigma_\gamma^2\mathbb{I}$. Then, from (13), we obtain

$$\begin{aligned} H_t &= (t-1)\Sigma_n^{-1} - (t-2)\Sigma_n^{-1}M^{-1}\Sigma_n^{-1} + \Sigma_\gamma^{-1} \\ &= \Big(\frac{(t-1)}{\sigma_n^2} - \frac{(t-2)\sigma_n^2\sigma_c^2}{\sigma_n^4\underbrace{(\sigma_c^2+\sigma_n^2)}_{:=f}} + \frac{1}{\sigma_\gamma^2}\Big)\mathbb{I} \\ &= \frac{(t-1)\sigma_\gamma^2 - (t-2)\sigma_c^2\sigma_\gamma^2/f + \sigma_n^2}{\sigma_n^2\sigma_\gamma^2}\mathbb{I}. \end{aligned} \tag{A94}$$

This implies

$$H_t^{-1} = \frac{\sigma_n^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2 - (t-2)\sigma_c^2\sigma_\gamma^2/f + \sigma_n^2}\mathbb{I}$$

whereby we obtain

$$\begin{aligned} R_t &= \frac{f\Big((t-1)\sigma_\gamma^2 + \sigma_n^2 + \sigma_c^2\Big)}{\sigma_c^2((t-1)\sigma_\gamma^2\sigma_n^2 + \sigma_c^2\sigma_\gamma^2 + f\sigma_n^2)}\mathbb{I} \quad \text{and} \\ R_t^{-1} &= \frac{\sigma_c^2((t-1)\sigma_\gamma^2\sigma_n^2 + \sigma_c^2\sigma_\gamma^2 + f\sigma_n^2)}{f\Big((t-1)\sigma_\gamma^2 + \sigma_n^2 + \sigma_c^2\Big)}\mathbb{I} = b_t\mathbb{I}. \end{aligned} \tag{A95}$$

Noting that $M_t = \frac{f}{\sigma_n^2\sigma_c^2}\mathbb{I}$ we then have

$$R_t^{-1}M_t = \frac{(t-1)\sigma_\gamma^2\sigma_n^2 + \sigma_c^2\sigma_\gamma^2 + f\sigma_n^2}{\sigma_n^2\Big((t-1)\sigma_\gamma^2 + \sigma_n^2 + \sigma_c^2\Big)}\mathbb{I} = \frac{(t-1)\sigma_\gamma^2\sigma_n^2 + \sigma_c^2\sigma_\gamma^2 + f\sigma_n^2}{(t-1)\sigma_\gamma^2\sigma_n^2 + f\sigma_n^2}\mathbb{I} = (1 + \frac{\sigma_c^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2\sigma_n^2 + f\sigma_n^2})\mathbb{I}. \tag{A96}$$

Subsequently, we obtain that for $t > 1$,

$$\log\Big(\det(R_t^{-1}M_t)\Big) = d\log\Big(1 + \frac{\sigma_c^2\sigma_\gamma^2}{(t-1)\sigma_\gamma^2\sigma_n^2 + f\sigma_n^2}\Big) \leq d\log\Big(1 + \frac{\sigma_c^2/\sigma_n^2}{(t-1)}\Big) \leq \frac{d\sigma_c^2/\sigma_n^2}{t-1}, \tag{A97}$$

whereby

$$\sum_{t=1}^T \log\Big(\det(R_t^{-1}M_t)\Big) \leq d\log\Big(1 + \frac{\sigma_c^2\sigma_\gamma^2}{f\sigma_n^2}\Big) + \sum_{t>1}^T \frac{d\sigma_c^2/\sigma_n^2}{t-1} \tag{A98}$$

$$\leq \frac{d\sigma_c^2}{\sigma_n^2}\Big(\frac{\sigma_\gamma^2}{f} + \log(T-1)\Big) \tag{A99}$$

where the last inequality follows since $\log(1+x) \leq x$ and $\sum_{s=1}^{T} \frac{1}{s} \leq \log(T)$.

Appendix D. Details on Experiments

In this section, we present details on the baselines implemented for stochastic CBs with unobserved true contexts.

Appendix D.1. Gaussian Bandits

For Gaussian bandits, we implemented the baselines as explained below.

TS_naive: This algorithm implements the following action policy at each iteration t,

$$a_t = \arg\max_{a \in \mathcal{A}} \phi(a, \hat{c}_t)^\top \theta_t,$$

where θ_t is sampled from a Gaussian distribution $\mathcal{N}(\mu_{t-1,\text{naive}}, \Sigma^{-1}_{t-1,\text{naive}})$ with

$$\Sigma_{t-1,\text{naive}} = \frac{\mathbb{I}}{\lambda} + \frac{1}{\sigma^2} \sum_{\tau=1}^{t-1} \phi(a_\tau, \hat{c}_\tau) \phi(a_\tau, \hat{c}_\tau)^\top$$
$$\mu_{t-1,\text{naive}} = \frac{\Sigma^{-1}_{t-1,\text{naive}}}{\sigma^2} \Big(\sum_{\tau=1}^{t-1} r_\tau \phi(a_\tau, \hat{c}_\tau) \Big).$$

TS_oracle: In this baseline, the agent has knowledge of the true predictive distribution $P(c_t|\hat{c}_t, \gamma^*)$. Consequently, at each iteration t, the algorithm chooses action

$$a_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t|\gamma^*)^\top \theta_t,$$

where θ_t is sampled from a Gaussian distribution $\mathcal{N}(\mu_{t-1,\text{poc}}, \Sigma^{-1}_{t-1,\text{poc}})$ with

$$\Sigma_{t-1,\text{poc}} = \frac{\mathbb{I}}{\lambda} + \frac{1}{\sigma^2} \sum_{\tau=1}^{t-1} \psi(a_\tau, \hat{c}_\tau|\gamma^*) \psi(a_\tau, \hat{c}_\tau|\gamma^*)^\top$$
$$\mu_{t-1,\text{poc}} = \frac{\Sigma^{-1}_{t-1,\text{poc}}}{\sigma^2} \Big(\sum_{\tau=1}^{t-1} r_\tau \psi(a_\tau, \hat{c}_\tau|\gamma^*) \Big).$$

For Gaussian bandits, the following figure shows additional experiment comparing the performance of our proposed Algorithm 1 for varying values of the number K of actions. All parameters are set as in Figure 1 (Left).

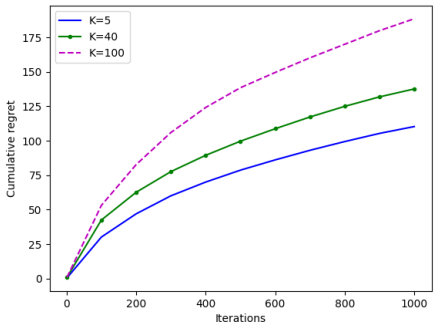

Figure A1. Bayesian cumulative regret of Algorithm 1 as a function of iterations over varying number K of actions.

Appendix D.2. Logistic Bandits

In the case of logistic bandits, we implemented the baselines as explained below.

TS_naive: This algorithm considers the following sampling distribution:

$$Q(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}}) \propto P(\theta^*) \prod_{\tau=1}^{t-1} \text{Ber}(\mu(\phi(a_\tau, \hat{c}_\tau)^\top \theta^*)).$$

However, due to the non-conjugateness of Gaussian prior $P(\theta^*)$ and Bernoulli reward likelihood, sampling from the above posterior distribution is not straightforward. Consequently, we adopt the Langevin Monte Carlo (LMC) sampling approach from [20]. To sample θ_t at iteration t, we run LMC for $I = 50$ iterations with learning rate $\eta_t = 0.2/t$ and inverse temperature parameter $\beta^{-1} = 0.001$. Then, θ_t is chosen as the output of the LMC after $I = 50$ iterations. Using the sampled θ_t, the algorithm then chooses the action a_t as

$$a_t = \arg\max_{a \in \mathcal{A}} \phi(a, \hat{c}_t)^\top \theta_t.$$

TS_oracle: This algorithm considers the following sampling distribution:

$$Q(\theta^*|\mathcal{H}_{t-1,r,a,\hat{c}}) \propto P(\theta^*) \prod_{\tau=1}^{t-1} \text{Ber}(\mu(\psi(a_\tau, \hat{c}_\tau|\gamma^*)^\top \theta^*)),$$

where $\psi(a_t, \hat{c}_t|\gamma^*) = \mathbb{E}_{P(c_t|\hat{c}_t,\gamma^*)}[\phi(a_t, c_t)]$ is the expected feature map under the posterior predictive distribution with known γ^*. As before, to sample from the above distribution, we use $I = 50$ iterations of LMC with learning rate $\eta_t = 0.2/t$ and inverse temperature parameter $\beta^{-1} = 0.001$. Using the sampled θ_t, the algorithm then chooses the action a_t as

$$a_t = \arg\max_{a \in \mathcal{A}} \psi(a, \hat{c}_t|\gamma^*)^\top \theta_t.$$

References

1. Srivastava, V.; Reverdy, P.; Leonard, N.E. Surveillance in an abruptly changing world via multiarmed bandits. In Proceedings of the IEEE Conference on Decision and Control (CDC), Los Angeles, CA, USA, 15–17 December 2014; pp. 692–697.
2. Aziz, M.; Kaufmann, E.; Riviere, M.K. On multi-armed bandit designs for dose-finding clinical trials. *J. Mach. Learn. Res.* **2021**, *22*, 686–723.
3. Anandkumar, A.; Michael, N.; Tang, A.K.; Swami, A. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE J. Sel. Areas Commun.* **2011**, *29*, 731–745. [CrossRef]
4. Srivastava, V.; Reverdy, P.; Leonard, N.E. On optimal foraging and multi-armed bandits. In Proceedings of the Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–4 October 2013; pp. 494–499.
5. Bubeck, S.; Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv* **2012**, arXiv:1204.5721.
6. Abe, N.; Biermann, A.W.; Long, P.M. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica* **2003**, *37*, 263–293. [CrossRef]
7. Agarwal, D.; Chen, B.C.; Elango, P.; Motgi, N.; Park, S.T.; Ramakrishnan, R.; Roy, S.; Zachariah, J. Online models for content optimization. *Adv. Neural Inf. Process. Syst.* **2008**, *21* , 17–24.
8. Auer, P.; Cesa-Bianchi, N.; Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **2002**, *47*, 235–256. [CrossRef]
9. Chu, W.; Li, L.; Reyzin, L.; Schapire, R. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 208–214.
10. Agrawal, S.; Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 127–135.
11. Lamprier, S.; Gisselbrecht, T.; Gallinari, P. Profile-based bandit with unknown profiles. *J. Mach. Learn. Res.* **2018**, *19*, 2060–2099.
12. Kirschner, J.; Krause, A. Stochastic bandits with context distributions. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14113–14122.
13. Yang, L.; Yang, J.; Ren, S. Multi-feedback bandit learning with probabilistic contexts. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main Track, Yokohama, Japan, 11–17 July 2020.
14. Kim, J.h.; Yun, S.Y.; Jeong, M.; Nam, J.; Shin, J.; Combes, R. Contextual Linear Bandits under Noisy Features: Towards Bayesian Oracles. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Valencia, Spain, 25–27 April 2023; pp. 1624–1645.
15. Guo, Y.; Murphy, S. Online learning in bandits with predicted context. *arXiv* **2023**, arXiv:2307.13916.
16. Roy, D.; Dutta, M. A systematic review and research perspective on recommender systems. *J. Big Data* **2022**, *9*, 59. [CrossRef]
17. Russo, D.; Van Roy, B. Learning to optimize via posterior sampling. *Math. Oper. Res.* **2014**, *39*, 1221–1243. [CrossRef]

18. Park, H.; Faradonbeh, M.K.S. Analysis of Thompson sampling for partially observable contextual multi-armed bandits. *IEEE Control Syst. Lett.* **2021**, *6*, 2150–2155. [CrossRef]
19. Neu, G.; Olkhovskaia, I.; Papini, M.; Schwartz, L. Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9486–9498.
20. Xu, P.; Zheng, H.; Mazumdar, E.V.; Azizzadenesheli, K.; Anandkumar, A. Langevin monte carlo for contextual bandits. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 24830–24850.
21. Harper, F.M.; Konstan, J.A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. (TIIS)* **2015**, *5*, 1–19. [CrossRef]
22. Hong, J.; Kveton, B.; Zaheer, M.; Ghavamzadeh, M. Hierarchical bayesian bandits. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Virtual, 28–30 March 2022; pp. 7724–7741.
23. Hong, J.; Kveton, B.; Katariya, S.; Zaheer, M.; Ghavamzadeh, M. Deep hierarchy in bandits. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 8833–8851.
24. Lattimore, T.; Szepesvári, C. *Bandit Algorithms*; Cambridge University Press: Cambridge, UK, 2020.
25. Russo, D.; Van Roy, B. An information-theoretic analysis of thompson sampling. *J. Mach. Learn. Res.* **2016**, *17*, 2442–2471.