Overcoming Prior Misspecification in Online Learning to Rank

Javad Azizi
Uni. of Southern California

Ofer Meshi
Google

Masrour Zoghi
Google

Maryam Karimzadehgan
Google

Abstract

The recent literature on online learning to rank (LTR) has established the utility of prior knowledge to Bayesian ranking bandit algorithms. However, a major limitation of existing work is the requirement for the prior used by the algorithm to match the true prior. In this paper, we propose and analyze adaptive algorithms that address this issue and additionally extend these results to the linear and generalized linear models. We also consider scalar relevance feedback on top of click feedback. Moreover, we demonstrate the efficacy of our algorithms using both synthetic and real-world experiments.

1 INTRODUCTION

Learning to rank (LTR) is the problem of ranking a set of items such that the resulting ranked list maximizes a utility function such as user satisfaction. This could be the ranking of search results for information retrieval (Liu et al., 2009), ranking the items in a recommendation system to increase user satisfaction (Karatzoglou et al., 2013; Falk, 2019), or ranking the ads in an ad placement system to enhance user engagement (Tagami et al., 2013).

In *offline* LTR, it is assumed that the *ground truth* utility of the lists has been provided and the goal is to learn a *scoring function* which can be used to rank the items (Zamani et al., 2017; Mitra and Craswell, 2017; Shen et al., 2018; Meng et al., 2020). In this setting, it is implicitly assumed that user behavior is time-invariant, however, in many real-world problems user preferences can change dynamically.

To address this issue, *online* LTR algorithms adaptively learn from user feedback. Various online learning algorithms have been designed for different user feedback models including the *cascade model* (Kveton et al., 2015; Zhong et al., 2021), the *position-based model* (Lagrée et al., 2016; Ermis et al., 2020a), as well as algorithms designed for multiple user models (Zoghi et al., 2017; Lattimore et al., 2018; Li et al., 2020). In this paper, we mainly focus on cascading bandits and defer extensions to other click models to future work.

Despite being adaptive, online LTR algorithms often suffer from the *cold-start* problem. More specifically, in the absence of any prior knowledge, the algorithm has to explore aggressively before it can start exploiting the information that it has learned about the users and items. The cost of this initial phase of exploration often renders online LTR algorithms impractical: this cost is particularly egregious in the ranking setting because of the large action space. One natural remedy for this problem is to infuse prior knowledge into the online LTR algorithm, which is the approach adopted by Kveton et al. (2022).

A major limitation of the algorithms and theoretical results in Kveton et al. (2022) is the assumption that the prior knowledge used by the algorithm does not deviate from the true prior. This assumption holds only in the most contrived of situations: for instance, for any time-varying data distribution, the prior obtained from previous observations cannot be a perfect prior for the future. This paper addresses this shortcoming by devising and analyzing algorithms that can make the most out of imperfect priors by adapting to online data.

More specifically, our contributions are as follows:

- In Section 3, we propose highly adaptive Gaussian Thompson Sampling (TS) algorithms for LTR in non-contextual and contextual settings. We consider the non-contextual setting in Section 3.1, the linear model in Section 3.2, and the generalized linear model (GLM) in Section 3.3.
- Our linear model in Section 3.2 handles scalar relevance feedback, generalizing our framework beyond click (binary) feedback.
- We derive Bayes regret bounds for our algorithms in Theorems 1 to 4. In the non-contextual setting, our bound is near-optimal. In the contextual settings, these bounds are the first Bayesian bounds in the literature to the best of our knowledge.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

- We conduct both synthetic and real-world experiments in Section 5. The synthetic experiments demonstrate that even though our algorithms start with an imperfect prior (e.g., flat), they quickly adapt to the environment and achieve competitive results compared to existing approaches. In particular, in the presence of prior misspecification, the performance of our algorithms does not deteriorate as severely as existing online LTR algorithms which use prior information. We conduct our real-world experiments on publicly available learning to rank datasets to test the efficacy of our algorithms in more realistic settings.

2 SETTING

We use the notation $[T] := \{1, \cdots, T\}$ for any integer T. For any vector or set v, let $v(i)$ be its i'th element.

We consider an online LTR problem where $\mathcal{L}$ is the set of items to choose from, with size L. The agent interacts with the environment, such as users in a recommender system, over T rounds. At round $t \in [T]$, the agent displays a ranked list A_t of $K \ll L$ items to the user, i.e., $A_t \in \Pi_K(\mathcal{L})$, where $\Pi_K(\mathcal{L})$ is the set of all tuples of K distinct items out of $\mathcal{L}$. The reward feedback depends on the attractiveness of items recommended to the user. We denote by $Y_{i,t}$ the attractiveness of item $i \in \mathcal{L}$ at round t, which is an independent Bernoulli random variable with mean $\nu_{i,t}$, i.e. $Y_{i,t} \sim \text{Bern}(\nu_{i,t})$.

Let $Y_t := Y_{A_t,t} := (Y_{A_t(k),t})_{k=1}^K$ be the vector of attraction indicators (feedback) at round t, and its corresponding reward be $R(Y_{A_t,t})$. In the *cascade model* (CM) we assume the user examines the first position in A_t with probability 1. If position $k \in [K]$ is examined and the item at that position is attractive, i.e., $Y_{A_t(k),t} = 1$, the user clicks on it and does not examine the rest. If the user does not click on the item at position k, they examine the next position, $k+1$. In this model, the reward at round t is 1 if there is any attractive item in A_t and 0 otherwise: $R(Y_{A,t}) = 1 - \left(\prod_{k=1}^K (1 - Y_{A_t(k),t})\right)$. Due to the independence of the attractiveness of items the expected reward is

$$r_{A,t} := \mathbf{E}[R(Y_{A,t})] = 1 - \prod_{k=1}^{K} (1 - \nu_{A(k),t}) \ . \quad (1)$$

We set $k_t = \min\{k \in [K] : Y_{A_t(k),t} = 1\}$ to be the click position, and use the convention $\min \emptyset = K$. That is if the user does not click on any of the items then $k_t = K$. Then $O_{i,t} = \sum_{k=1}^{k_t} \mathbf{1}(i = A_t(k))$ is the indicator of item i being examined (observed) at round t. The set of rounds where item i is observed by round t is $\mathcal{T}_{i,t} := \{s \le t : i \in A_s, i \le k_s\}$, with size $T_{i,t} := |\mathcal{T}_{i,t}|$.

Let $A_t^* = \arg\max_{A \in \Pi_k(L)} r_{A,t}$ be the best ordered list (best action) at round t. Then the *frequentist regret* at round t is $\mathcal{R}_t := r_{A_t^*} - r_{A_t}$, and the cumulative (frequentist) regret for horizon T is $\mathcal{R}(T) := \mathbf{E}[\sum_{t=1}^T \mathcal{R}_t]$. We define the *Bayes regret* as the expectation of the frequentist regret over the problem instances of the attractiveness probabilities ν:

$$\mathcal{BR}(T) := \mathbf{E}[\mathcal{R}(T)] = \mathbf{E}\left[\sum_{t=1}^{T} \mathbf{E}_t[\mathcal{R}_t]\right] ,$$

where $\mathbf{E}_t[\cdot] := \mathbf{E}[\cdot|\mathcal{H}_t]$ is the conditional expectation given $\mathcal{H}_t$, the trajectory of clicks and actions up to but not including round t. The second equality above holds by the tower rule.

3 THOMPSON SAMPLING FOR LTR

In this section, we develop TS algorithms for LTR under various parameterizations, including the non-contextual setting as in Section 2, and also its extension to linear and generalized linear contextual bandits in Sections 3.2 and 3.3. We assume the prior is such that $\nu \sim P_0(\mu)$, where μ is the prior parameter. We denote the posterior at round t with $P_t(\hat{\nu}_t)$, where $\hat{\nu}_t$ is the posterior estimate of ν given $\mathcal{H}_t$.

3.1 Gaussian Thompson Sampling

In this section, we propose an adaptive Gaussian TS algorithm for the LTR problem under the cascade model. Agrawal and Goyal (2013) showed that Gaussian TS performs very well when applied to bandits with a bounded distribution. Inspired by this observation, we propose a Gaussian TS algorithm for our LTR problems which has Bernoulli (so bounded) rewards, i.e., implicitly assume a Gaussian feedback while it is Bernoulli. See Gaussian TS paragraph of Section 4 for further details.

The pseudo-code of our algorithm is in Algorithm 1. We begin with a Gaussian prior $P_0(\mu) = \mathcal{N}(\mu_0, \sigma_0)$, with mean μ_0 and standard deviation σ_0 (Line 1). The posterior updates are as follows (Line 3):

$$Y_{i,t} \mid \nu_i \sim \mathcal{N}(\nu_i, \sigma^2) \qquad \text{(Likelihood)}$$
$$\nu_{i,t} \sim P_{i,t}(\nu_i|\mathcal{H}_t) = \mathcal{N}(\hat{\nu}_{i,t}, \hat{\sigma}_{i,t}^2) \qquad \text{(Posterior)} , \quad (2)$$

where by Murphy (2007, Eqs 20 and 24)

$$\hat{\nu}_{i,t} = \left(\frac{\sum_{\ell=1}^t Y_{i,\ell} O_{i,t}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \hat{\sigma}_{i,t}^2 \text{ and } \hat{\sigma}_{i,t}^2 = \left(\frac{1}{\sigma_0^2} + \frac{T_{i,t}}{\sigma^2}\right)^{-1}$$

Note that the posterior at round $t+1$ matches the one computed in Eq. (2) in a Gaussian TS (see Murphy (2007) for detail). We compare this to Beta TS (Kveton et al., 2022) in Appendix A.1, where we highlight its advantages over the Beta TS.

At round t, our algorithm computes the posterior based on $\mathcal{H}_t$ as given in Eq. (2). Next, it acquires a sample from the

Algorithm 1 TS for LTR

1: **Initialize:** $\mathcal{H}_1 = \emptyset$ and set prior $P_0 := \mathcal{N}(\mu_0, \sigma_0)$
2: **for** $t = 1, \dots, T$ **do**
3: Compute posterior $P_{i,t}(\nu_i|\mathcal{H}_t)$ based on Eq. (2)
4: Sample $\tilde{\nu}_{i,t} \sim P_{i,t}(\nu|\mathcal{H}_t)$ for all $i \in \mathcal{L}$ # Posterior sample
5: $A_t \in \arg\max_{\substack{A \subset \mathcal{L} \\ |A|=K}} \sum_{i \in A} \tilde{\nu}_{i,t}$ # Recommend the top posterior samples
6: Observe Y_t and let $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{(A_t, Y_t)\}$
7: **end for**

posterior distribution for each item and recommends the items corresponding to the top K samples (Lines 4 and 5). Finally, it updates $\mathcal{H}_{t+1}$ by including the action and reward pair (Line 6).

We analyze this algorithm with the default prior $\mathcal{N}(0, 1)$. In Section 5.1, we show empirically that the effect of this prior vanishes quickly. However, in a Bernoulli environment like our non-contextual model $\mathcal{N}(0, 0.25^2)$ might be more favorable given that Bernoulli random variables are 0.25-subgaussian.

Algorithm 1 Analysis: We follow a similar line of reasoning as Bubeck and Liu (2013, Theorem 1) for Gaussian TS Bayes regret. The challenge is to apply that result to the cascading bandits. We overcome this using a results about the cascade model shown in Kveton et al. (2022, Lemma 7), which bounds $\mathcal{R}_t$ with $\sum_{k \in A_t} r_{A_t^*(k),t} - r_{k,t}$.

Theorem 1 (Gaussian TS)**.** *The Bayes regret of Algorithm 1 with prior $\mathcal{N}(0, 1)$ satisfies*

$$\mathcal{BR}(T) \leq 22K\sqrt{LT} \ .$$

The proof is in Appendix A. Compared to the result of Kveton et al. (2022), it appears that we need to pay a $O(\sqrt{K})$ price for not knowing the prior. If we use the true prior we can shave off this extra term by using exact confidence bounds as used in Kveton et al. (2022). In terms of L and T, this bound is of optimal order since they match the lower bound for Gaussian TS (Agrawal and Goyal, 2013). In addition, the constants are much smaller compared to the frequentist bounds of Zhong et al. (2021) (see Section 4). Recently, Vial et al. (2022) proved a matching lower bound of $\Omega(\sqrt{LT})$, which shows that our algorithm achieves near-optimal regret bound.

In the next two sections, we deal with the contextual setting. In this setting, we use a feature vector $x_{i,t} \in \mathcal{X}_t \subset \mathbb{R}^d$ for each item i at round t, where $\mathcal{X}_t$ is the set of feature vectors for the items to be ranked at round t. For instance, in an information retrieval system, this vector encodes user, query, and item information. We use the notation $x_t := (x_{i,t})_{i \in \mathcal{L}}$ for the array of feature vectors in the list.

3.2 Linear Contextual LTR

In this setting, we assume that the reward generated for each item is governed by a linear function. This setting is relevant to applications where the feedback (Y) is not binary (not Bernoulli clicks) but a real number such as the amount of time the user spent on a page or their level of satisfaction with the page, etc. We continue to assume a cascade user model in the sense that the user scans down the list and either generates no reward on any item in the list or generates independent rewards for some items and then abandons the list at one item.

The reward at round t for action A is the sum of item level rewards of the observed items, i.e.,

$$R(Y_{A,t}) = \sum_{k=1}^{K} O_{A(k),t} Y_{A(k),t} \ .$$

Recall from Section 2 that $O_{i,t}$ is the independent random variable that indicates whether item i was observed at round t. By independence we have $r_{A,t} = \sum_{k=1}^{K} r_{A(k),t}$. Note that this reward is the same as the reward in the case of the cascade model if Y is binary since the right-hand side of Eq. (1) is 0 when no item in the list is clicked and 1 when there is a click. Also, in the cascade model there is at most one click.

We begin by laying out our assumptions in this setting.

Assumption 1 (Reward)**.** *If we have $O_{i,t} = 1$, then the feedback for item t is $Y_{i,t} = x_{i,t}^\top \theta + \eta_t$, where θ is the unknown parameter of the problem and η_t is σ^2-sub-Gaussian. Thus, $r_{i,t} = x_{i,t}^\top \theta$ is the mean reward (feedback) of item i at round t given feature $x_{i,t}$.*

Assumption 2 (Features)**.** *We assume $\|x_{i,t}\| \leq 1$ for all i, t, and that the set of feature vectors $\mathcal{X} \subset \mathbb{R}^d$ is compact.*[1]

Assumption 3 (Parameters)**.** *We assume $\|\theta\| \leq S$ for some $S \in \mathbb{R}^+$.*

Algorithm 2 contains the pseudo-code of our linear TS algorithm. In Algorithm 2 we have used $\beta_t(\delta_t) := \sigma^2\sqrt{2\log\frac{(\lambda+t)^{d/2}\lambda^{-d/2}}{\delta_t}} + \sqrt{\lambda}S$, where $\delta_t = \delta/2^{\max(1, \lceil \log(t) \rceil)}$ and $\delta = \frac{1}{T(\log T + 2)}$. This algorithm is similar to Algorithm 1, with the main difference being the calculation of the posterior sample. In particular, we use *regularized least-squares* estimate of θ (Line 5) and sample $\tilde{\theta}_t$ from the posterior (Line 6). Then, we generalize the posterior sample for item i with $x_{i,t}^\top \tilde{\theta}_t$ in Line 7.

Algorithm 2 Analysis: We derive the first Bayesian regret analysis for linear TS for cascading bandits. Our innovation lies in providing a Bayes regret by converting the frequentist analysis in prior work on linear TS (Abeille and Lazaric, 2017) to the cascade model and bound the Bayes regret.

[1] We use $\|\cdot\|$ to denote the ℓ_2 norm.

Algorithm 2 Linear TS for LTR

```
Input: The regularization parameter λ ∈ ℝ⁺, S
Initialize: V_1 ← λI_d, r̄_1 ← 0
for t = 1, ..., T do
    Receive the current context x_t
    Let θ̂_t = V_t^{-1} r̄_t.   # Compute the posterior parameter
    Sample θ̃_t ~ N(θ̂_t, β_t^2(δ_t) V_t^{-1})   # Posterior sample
    Recommend A_t ∈ argmax_{A⊂L, |A|=K} Σ_{i∈A} x_{i,t}^T θ̃_t   # Recommend the top posterior samples
    Observe Y_t
    V_{t+1} ← V_t + Σ_{i∈A_t} O_{i,t} x_{i,t} x_{i,t}^T
    r̄_{t+1} ← r̄_t + Σ_{i∈A_t} O_{i,t} x_{i,t} Y_{A_t(i),t}
end for
```

Theorem 2 (Algorithm 2 Regret)**.** *Under Assumptions 1 to 3, the Bayes regret of Algorithm 2 is* $\mathcal{BR}(T) = \tilde{O}\left(d^{3/2}K\sqrt{T}\right)$*, where* $\tilde{O}$ *hides logarithmic terms.*

Note that compared to Theorem 1, the dependence on L is replaced with a d dependence.[2] Compared to frequentist upper bound in Vial et al. (2022), i.e., $\tilde{O}(\sqrt{Td(d+K)})$, our regret bound is order-wise sub-optimal, but it does not include huge constants like theirs. To the best of our knowledge, no lower bound is known for cascading bandits in the linear case. We prove Theorem 2 in Appendix B starting with a decomposition that separates the posterior sample error and the estimation error. Then, we bound each term using ellipsoidal high probability bounds from Abbasi-Yadkori et al. (2011); Abeille and Lazaric (2017).

3.3 Generalized Linear Contextual LTR

In a click model, it is natural to consider the clicks as Bernoulli random variables as in Section 2. In this section, we employ a *generalized linear model* (e.g. *logistic* model) to model the Bernoulli attractiveness random variables in a contextual setting.

We model the attractiveness as $Y_{i,t} \sim \text{Bern}(\mu(x_{i,t}^\top \theta))$ where $\mu(x)$ is the so-called *link* function, e.g., $\mu(x) = 1/(1+\exp(-x))$ is the Sigmoid function. The reward functions, R and r, are defined the same as in Eq. (1) where ν is replaced with $\mu(x^\top\theta)$, which is the probability of attractiveness for item i at round t. We have the following assumption on the reward noise.

Assumption 4 (Reward)**.** *We assume that if* $O_{i,t} = 1$ *then the feedback for item t is* $Y_{i,t} = \mu(x_{i,t}^\top\theta) + \eta_t$*, where* η_t *is* σ^2*-sub-Gaussian.*

The pseudo-code of our algorithm for this setting is in Algorithm 3.

Algorithm 3 GLM TS for LTR

```
1: Input: The regularization parameter λ ∈ ℝ⁺, S, and K
2: Initialize: V_1 ← λI_d.
3: for t = 1, ..., T do
4:     Receive the current context x_t.
5:     Solve Eq. (3) for θ̂_t   # Compute the posterior parameter
6:     Sample θ̃_t ~ N(θ̂_t, β_t^2(δ_t) V_t^{-1} / κ^2)   # Laplace posterior sample
7:     A_t ∈ argmax_{a⊂L, |a|=K} Σ_{i∈a} x_{i,t}^T θ̃_t   # Recommend the top posterior samples
8:     Observe Y_t
9:     V_{t+1} ← V_t + Σ_{i∈A_t} O_{i,t} μ(θ̂_t^T x_{i,t})(1 − μ(θ̂_t^T x_{i,t})) x_{i,t} x_{i,t}^T
10: end for
```

We define $\kappa := \inf_{\theta\in\Theta, x\in\mathcal{X}} \dot{\mu}(x^\top\theta)$ where $\dot{\mu}(z) = \frac{\partial\mu(z)}{\partial z}$. The general structure of the algorithm is the same as before. Therefore, we discuss only the new parts. At round t, we estimate θ (Line 5) by solving the following equation

$$\sum_{s=1}^{t}\sum_{k=1}^{K} O_{k,t}\left(Y_{A_s(k),s} - \mu(x_{A_s(k),s}^\top\hat{\theta}_t)\right)x_{A_s(k),s} = 0\,, \tag{3}$$

for $\hat{\theta}_t$, using a *iteratively reweighted least square* (IRLS) oracle (Green, 1984). Next, we sample from the Laplace approximation of the posterior (Bishop and Nasrabadi, 2006, Chapter 4.5.1) in Line 6. We provide an alternative algorithm in Appendix D which takes single Newton steps (Gentile and Orabona, 2012) and therefore it is more computationally attractive compared to using IRLS.

Algorithm 3 Analysis: We add the following assumption on the properties of μ, and derive a regret bound independent of L as follows.

Assumption 5 (Link Function)**.** *The link function,* $\mu : \mathbb{R} \mapsto \mathbb{R}$*, is continuously differentiable, Lipschitz with constant* ℓ *and* $\kappa > 0$.

For example, $\ell = 1/4$ for the Sigmoid function, and κ depends on $\sup_{\theta\in\Theta, x\in\mathcal{X}} x^\top\theta$, which by Assumptions 2 and 3 is at most S.

Theorem 3 (Algorithm 3 Regret)**.** *The Bayes regret of Algorithm 3 under Assumptions 2 to 4 satisfies*

$$\mathcal{BR}(T) \le K\frac{\ell}{\kappa}\big(\beta_t(\delta') + \gamma_t(\delta')d\big)\sqrt{2Td\log(1+\frac{T}{\lambda})} + K\frac{2\gamma_t(\delta')}{0.1\kappa}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}$$

with probability at least $1-\delta$ *where* $\delta' = \frac{\delta}{4T}$*. Here* $\gamma_t(\delta) = \beta_t(\delta)\sqrt{cd\log(c'd/\delta)}$ *with* c, c' *constants such*

[2] The O operator does not hide any T term.

that $\mathbf{P}\left(\|Y_{i,t} - x_{i,t}^\top\theta\| \leq \sqrt{cd\log(c'd/\delta)}\right) \geq 1-\delta$. *We also show by choosing a proper* δ *and* γ *the regret is* $\tilde{O}\left(K\frac{\ell}{\kappa}\sigma^2 d\sqrt{Td}\right)$.

We prove this in Appendix C, using an extension of the linear TS regret bounds and the definition of κ and ℓ. To the best of our knowledge, there is no prior work on cascading bandits with comparable Bayes regret bound. Santara et al. (2022) developed an algorithm for a routing problem in a cascading bandit setting. Their regret bound is $O\left(d\log(1+T)\sqrt{Te^{2S}(K+d\log(1+\frac{KdT}{\delta})}\right) = \tilde{O}\left(dTe^S\sqrt{T(K+d)}\right)$ for $\delta \in (0,1/e]$ with probability $1-\delta$. This is better than our bound by a factor of $\sqrt{K}$, but it has an exponential dependence on S which could be as bad as $1/\kappa$ (see Remark 1 in Santara et al. (2022)).

In prior works on GLM bandits, $1/\kappa$ usually appears in the regret bounds. However, depending on $\mathcal{X}$, this term could be arbitrarily large. As such, it is desirable if our regret analysis does not contain κ. We employ an information-theoretic approach as in Neu et al. (2022) to derive the following result for our LTR setting. We define a new information ratio and bound it for each item in the recommendation list, which readily gives an upper bound on the Bayes regret of our algorithm. We focus on the Sigmoid function for the simplicity of the proofs.

Theorem 4 (Algorithm 3 Regret without κ)**.** *If Assumptions 2 to 4 hold and* μ *is the Sigmoid function, then the Bayes regret of Algorithm 3 satisfies*

$$\mathcal{BR}(T) \leq K\sqrt{2L(d\log(2ST+1)+1)} = \tilde{O}(K\sqrt{dLT})$$

Our information-theoretic proof of this result is outlined in Appendix C.1. We note that the dependence on L is not desirable in a contextual setting, but this bound is based on the state-of-the-art results for GLM bandits as it avoids κ.

4 RELATED WORK

Cascading bandits: The cascade click model was introduced in Richardson et al. (2007); Craswell et al. (2008) and applied to online LTR in Kveton et al. (2015). Zhong et al. (2021) proposed `TS-Cascade` for non-contextual cascading bandits which is a semi-Thompson sampling (TS) algorithm (dependent posterior samples used) with Gaussian updates which start with an uninformative Gaussian prior (flat prior). They present a frequentist regret upper bound for this algorithm, however, their regret bound includes a huge constant (Lemma 4.3 and the last equation in Section 4 therein). Our Algorithm 1 improves on their algorithm by applying an exact TS Gaussain update. Our analysis also improves the regret bound, although it is a Bayes regret.

Gaussian TS: Agrawal and Goyal (2013) showed the problem-independent regret bound of Gaussian TS for a $[0,1]$ reward K-armed bandit problem is $O(\sqrt{KT\log K})$ for $T > K$. Bubeck and Liu (2013) used refined confidence bounds and improved this bound by removing the $\log K$ factor. Abeille and Lazaric (2017) revisited linear TS for contextual bandits under the frequentist setting. In a related work, Liu et al. (2022) quantified the amount of miss-specification in Gaussian TS applied to Bernoulli bandits. LTR click models usually assume Bernoulli click random variables which is $[0,1]$-bounded. Similarly, we introduce a Gaussian TS for our LTR problem. We propose Algorithm 2 and the first Bayesian regret analysis for Linear TS in the cascading model. Zhong et al. (2021) proposed `LinTS-Cascade` for linear cascade model and showed $O(\min\{\sqrt{d},\sqrt{\log L}\}Kd\sqrt{T}(\log T)^{3/2})$ frequentist regret where the constant is $(1+\lambda)(1+8\sqrt{\pi}e^{7/(2\lambda^2)})$ which could be very large. Faury et al. (2020) proposed the state-of-the-art UCB algorithm for logistic bandits which achieves $O(\sqrt{\kappa T}d\log T)$ regret bound. This result improves on the classic result of Filippi et al. (2010) removing a $\sqrt{\kappa}$ factor. Depending on $\mathcal{X}$, this number could be arbitrarily large. Neu et al. (2022) provide a regret bound for TS over logistic bandits which is $\tilde{O}(\sqrt{dLT})$ for L arms. This bound is $\sqrt{L}$ worse than the above but it is κ free which makes it viable in that sense.

Related algorithms: The state-of-the-art prior works comparable to our work include (i) prior-free online LTR algorithms like `CascadeKL-UCB`, `CascadeUCB1` from Kveton et al. (2015), and `TopRank` (Lattimore et al., 2018), (ii) the only prior-based adaptive algorithms, `BayesUCB` and `TS` (Kveton et al., 2022), and finally, (3) non-adaptive prior-based algorithms where the items are ranked according to their maximum-a-posteriori (MAP) attraction probabilities estimated from the same prior as in `BayesUCB` and `TS`, such as `Ensemble` in (Kveton et al., 2022).

`CascadeLinUCB` is a variant of the LinUCB algorithm applied to the linear cascade bandits setting by (Zong et al., 2016). Note that `CascadeLinTS` from that paper is almost the same as our `LinTS-LTR`, except the calculation of the variance of posterior. However, they do not have an analysis for this algorithm. In the linear setting, however, the state-of-the-art is `CascadeWOFUL` (Vial et al., 2022) which uses Bernstein and Chernoff confidence bounds to develop a variance-aware algorithm. `LinTS-Cascade` (Zhong et al., 2021) is again similar to our `LinTS-LTR` but with different variance calculations and it only has frequentist analysis with huge constants.

There are several other click models considered in the bandit literature. Recently, Ermis et al. (2020b) proposed a linear UCB and a linear TS algorithm for the *position based* click model (PBM). They, however, do not analyze the regret of these algorithms.

GLM bandits: Several recent works study GLM and especially logistic bandits. Dong et al. (2019) analyzed the performance of TS for logistic bandits using an information-theoretic approach which results in a regret bound without κ in it but with other potentially large constants. Neu et al. (2022) improved on this and proposed a regret bound free of large constants. In Theorem 4 we take a similar approach to this. Dumitrascu et al. (2018) proposed an improvement over the Laplace estimate of the posterior for logistic TS but did not bound its regret. Several other recent works including Faury et al. (2020); Abeille et al. (2021); Faury et al. (2022) study UCB and optimism in face of uncertainty (OFU) based algorithms which we could further consider for future work.

Offline initialization: Considering the cold-start issues, we can learn the prior information from the historical data and use it in near-optimal algorithms like in Kveton et al. (2022). The Web30K experiment in Kveton et al. (2022) practically does this.

5 EXPERIMENTS

In this section, we aim at evaluating our algorithms empirically and compare them to prior work. The goal is to validate the theory in practice using synthetic and real-world data. In particular, we are trying to answer the following research questions;

1. How does our non-contextual algorithm Algorithm 1 compare to prior work on non-contextual LTR? We investigate this in Fig. 1.
2. How does prior initialization and misspecification affect our non-contextual algorithm? Fig. 2 and further experiments of Figs. 8, 15 and 16 in Appendix F answer this.
3. For a linear or logistic reward model, how does Algorithms 2 and 3 compare to prior work in terms of regret? We conduct the corresponding experiments for the linear and logistic model in Figs. 3 and 4, respectively. Figs. 9 and 10 experiments in Appendix F explore this further.
4. How would our algorithms perform if the theoretical assumptions are violated? We explore this using two datasets in Figs. 5 and 6. Further experiments Figs. 11 to 14 are in Appendix F.

The default setting for our synthetic experiments is $L = 30$, $K = 3$, $T = 10000$, $S = 1$, and $\lambda = 1/10^4$. Our preliminary analysis shows that the few hyper-parameters (λ, S) of our algorithms do not change the behavior much and thus we did not include a tuning study. This is also in line with the regret bounds which show the effect of the hyperparameters is negligible. We compare the cumulative regret of the algorithms over 100 simulation replications. In the synthetic experiments, each replication is a new instance sampled from a fixed prior.

Our Algorithm 1 is shown as `GTS`, Algorithm 2 as `LinTS-LTR`, and Algorithm 3 with Sigmoid link function as `LogTS-LTR`. Here, `GTS-Pmean` is `GTS` with a prior mean initialized to the mean of the offline data (instead of $\mathcal{N}(0, 1)$ uninformative prior). `GTS-P` uses both the mean and variance of the offline data for the prior initialization.

5.1 Synthetic Experiment: Non-contextual Setting

In the non-contextual setting, our baselines are `TS` (shown as `TS-Beta` to emphasis beta updates versus Gaussian) and `BayesUCB` (shown as `BayesUCB-Beta`) from Kveton et al. (2022) which are the state-of-the-art and shown to outperform the other algorithms. These algorithms use the exact prior so we call them *prior-informed* algorithms. The rest of the algorithms here which do not use the prior are called *prior-agnostic*. For instance, we include `TS-Cascade` which is the closest algorithm to `GTS` in the literature from Zhong et al. (2021) and uses dependent posterior samples in a semi-TS algorithm (see Section 4).

For the non-contextual setting, in Fig. 1 (and Figs. 8, 15 and 16 in Appendix F), we let $(\beta_{1,i}, \beta_{2,i})$ be the Beta prior distribution parameters (used by `TS` and `BayesUCB`). We let $\beta_{2,i} = 10$ and sample $\beta_{1,i}$ from [10] uniformly at random, for all items i. We sample $(\beta_{1,i})_{i=1}^{L}$ for 20 times, and for each one sample the attraction probability of items as $\nu_i \sim \text{Beta}(\beta_{1,i}, \beta_{2,i})$ 20 times. Each algorithm is simulated on these 400 bandit instances. This procedure generates attraction probabilities between 0.09 and 0.5, so no item is overly attractive.

Fig. 1 shows the cumulative regret for all the algorithms on a cascade model. We can observe that `GTS` competes closely with `TS-Beta` and `BayesUCB-Beta`, despite not having access to the true prior. Also, `TS-Cascade` seems to underperform significantly. We conjecture this could be due to its dependent posterior samples which could pollute its posterior statistics (not MAP estimates). In Fig. 8 of Appendix F we removed `TS-Cascade` to zoom in on the difference between `GTS`'s. `GTS-Pmean` and `GTS-P` correct `GTS` at the beginning as they start with a better prior mean, the gain, however, is not significant which shows our TS algorithms are highly adaptive. Figs. 15 and 16 in Appendix F confirm this observation further. Also, `GTS-P` shows slightly worse performance which seems to highlight a drawback of initializing a Gaussian TS with a beta prior.

In Fig. 2, the true prior is $Beta(1, 10)$ but the prior fed

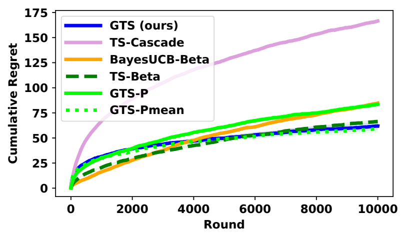

Figure 1: Synthetic non-contextual setting.

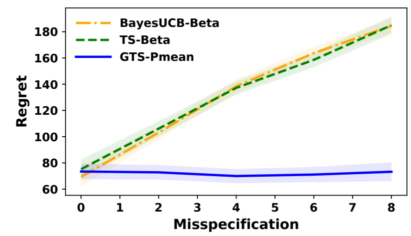

Figure 2: Synthetic non-contextual setting with prior misspecification.

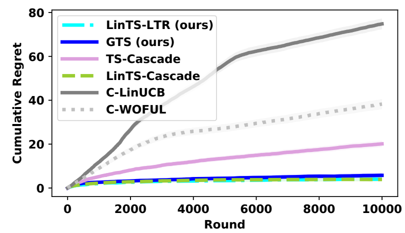

Figure 3: Linear setting experiment.

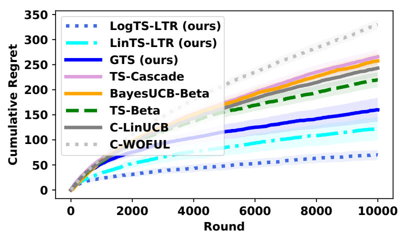

Figure 4: Logistic setting experiment.

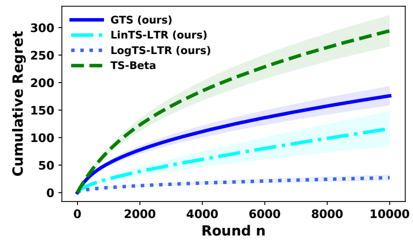

Figure 5: Web30k experiment with $K = 10$.

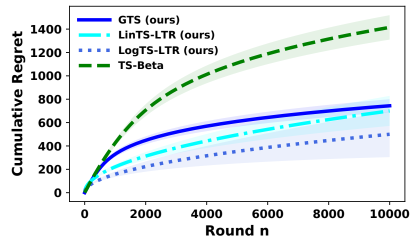

Figure 6: Istella experiment with $K = 10$.

to the algorithms is Beta$(1 + c, 10 - c)$, with $c \in [0, 8]$. Therefore, when $c > 0$, the prior is misspecified. We plot the final cumulative regret of the algorithms at $T = 1000$. This figure illustrates that our algorithm[3] is robust against the prior misspecification, while the performance of prior-dependant algorithms degrades, as expected.

5.2 Synthetic Experiment: Linear Setting

In the linear setting, the state-of-the-art baseline algorithm is `CascadeWOFUL` (Vial et al., 2022), labeled as `C-WOFUL`, which deploys a Bernstein UCB algorithm and adapts to the variance of the rewards. We removed `BayesUCB-Beta` in some experiments as it has almost the same performance as `TS-Beta`.

In this section, we sample $\theta \in [0, 1]^d$ and $x_{i,t} \in [0, 1]^d$ from a uniform distribution and normalize them to unit norm. This aligns with the experimental setup of the linear LTR algorithm in prior work (Zhong et al., 2021; Vial et al., 2022).

Fig. 3 demonstrate the cumulative regret of the state-of-the-art linear algorithm and our `LinTS-LTR`. It shows that our algorithm and `LinTS-Cascade` (Zhong et al., 2021) outperform the prior work. Fig. 9 in Appendix F separates the underperforming algorithms and illustrates that `LinTS-Cascade` and `LinTS-LTR` outperform all others including `GTS`, showcasing the benefit of including the contextual information.

[3]`GTS-Pmean` or `GTS` both perform similarly here so we only keep `GTS-Pmean`

5.3 Synthetic Experiment: Logistic Setting

For the logistic setting, as far as we know there are no comparable prior works for the cascade model. We include `CascadeLinUCB` (shown as `C-LinUCB`) for completeness in the experiments, although it is empirically shown to underperform `CascadeWOFUL`. In this section, we sample θ and x_i in the same way as in Section 5.2, but set $\nu_i = \mu(\theta^\top x_i)$. We set $\beta_{2,i} = 10$ and $\beta_{1,i} = 10\nu_i/(1 - \nu_i)$ so the mean of Beta prior of item i is ν_i.

We show the cumulative regret of our algorithms in Fig. 4, where our algorithm's superior performance is highlighted. A zoomed-in version of this in Fig. 10 in Appendix F shows `LogTS-LTR` outperforms other algorithms (including the prior-informed ones). Moreover, even when considering only the first 1000 steps, `LogTS-LTR` still outperforms the rest, showing that it is highly adaptive.

5.4 Real-world Experiments

In this section, we first experiment with the Microsoft Learning to Rank **Web30K** dataset[4] (Qin and Liu, 2013), which contains ∼18000 training and ∼6000 test queries, with an average of $L = 120$ documents per query. Each query and document pair has 136 features which we normalize and take as the context. We follow the setting in Kveton et al. (2022). In particular, priors are generated from this dataset by training $M = 10$ state-of-the-art deep-learning LTR model (Qin et al., 2021), each on

[4]https://www.microsoft.com/en-us/research/project/mslr/

a randomly selected training subset. This gives M scores for each query-document pair i, $(s_{i,j})_{j=1}^{M} \in [0,1]$. Then, $\text{Beta}(\sum_{j=1}^{m} s_{i,j}, \sum_{j=1}^{m}(1-s_{i,j}))$ is used as the Beta prior. See Kveton et al. (2022, Section 7.4) for more details.

In Fig. 5 we depict the regret of the most competitive algorithms in this dataset for 50 simulation replications on each query in the test dataset. We observe that `GTS` has superior performance compared to `TS-Beta`. Also, `LogTS-LTR` outperforms other algorithms by using the context and its nonlinear power. It is interesting to see how `LinTS-LTR` performs better than `GTS` but worse than `LogTS-LTR`, as expected. Fig. 11 in Appendix F shows a closer look of this result. Fig. 13 in Appendix F shows the performance of other algorithms that underperform the ones here for completeness.

Next, we experiment with **Istella**.[5] This is one of the largest publicly available learning-to-rank datasets, particularly useful for large-scale experiments on the efficiency and scalability of LETOR solutions. It is composed of 33,018 queries and 220 features representing each query-document pair. It consists of 10,454,629 examples (query-document) labeled with relevance judgments ranging from 0 (irrelevant) to 4 (perfectly relevant). The average number of per-query examples is $L = 316$. This dataset is split into train and test sets according to an 80%-20% scheme.

Fig. 6 shows the results of our Istella experiment. We illustrate the regret of the most competitive algorithms in this dataset for 50 simulation replications on each query in the test dataset. We can see that `LogTS-LTR` still outperforms `LinTS-LTR`, `GTS`, and `TS-Beta`, but by a smaller margin. We also observe the higher variance in this experiment compared to Web30K. Our conjecture is that this is due to the high sparsity of the features in this dataset. Fig. 12 in Appendix F shows a closer look of this result. Fig. 14 in Appendix F shows that `CascadeWOFUL` underperforms the other algorithms here.

In summary, the main finding in this section is that `LogTS-LTR` outperforms all other 4 algorithms (cf. Figs. 5 and 6). This suggests that these real-world settings could have common non-linear characteristics with the logistic setting (Fig. 4), even though we do not control their true reward.

6 CONCLUSION

We proposed highly adaptive Gaussian TS algorithms for online learning to rank that handle prior misspecification. Our algorithms exploit contextual information. We extend prior work on cascading bandit to new forms of relevance feedback other than clicks. In particular, our linear model handles scalar feedback. Our Bayes regret upper bounds are the first of their own sort, and in the non-contextual setting our bound is near-optimal. Finally, our extensive experiments on synthetic and real-world datasets demonstrate the efficacy of our algorithms.

Prior-dependent regret bounds are of interest for future work. We can achieve this type of bounds using the information-theoretic analysis in Liu et al. (2022); Lu and Van Roy (2019). Extending our work to other click models such as *dependant click model* (DCM) (Guo et al., 2009) in the contextual framework seems promising (Liu et al., 2018; Santara et al., 2022; Ermis et al., 2020b). Vial et al. (2022) developed a GLM bandits algorithm based on UCB for DCM and it would be interesting to see how our TS algorithm would compare to it in DCM. Providing a lower bound for the contextual setting is of interest as well.

Meta-learning the prior in a hierarchical structure (Hong et al., 2022) over the queries (cluster the queries) could potentially alleviate the cold-start problem too. Approximate posterior sampling algorithms (Dumitrascu et al., 2018) generalize our TS algorithm to a vast spectrum of feedback distribution beyond Gaussian and Bernoulli which we deal with here. Improved alternatives for posterior sampling such as Feel good TS (Zhang, 2022) and Maillard Sampling algorithm (Bian and Jun, 2022) have the potential to further improve the performance of our algorithms.

[5] http://blog.istella.it/istella-learning-to-rank-dataset/

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Abeille, M., Faury, L., and Calauzènes, C. (2021). Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR.

Abeille, M. and Lazaric, A. (2017). Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR.

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

Audibert, J.-Y. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836.

Bian, J. and Jun, K.-S. (2022). Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pages 54–72. PMLR.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Bubeck, S. and Liu, C.-Y. (2013). Prior-free and prior-dependent regret bounds for thompson sampling. *Advances in neural information processing systems*, 26.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94.

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

Dumitrascu, B., Feng, K., and Engelhardt, B. (2018). Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, 31.

Ermis, B., Ernst, P., Stein, Y., and Zappella, G. (2020a). Learning to rank in the position based model with bandit feedback. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2405–2412.

Ermis, B., Ernst, P., Stein, Y., and Zappella, G. (2020b). Learning to rank in the position based model with bandit feedback. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2405–2412.

Falk, K. (2019). *Practical recommender systems*. Simon and Schuster.

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

Faury, L., Abeille, M., Jun, K.-S., and Calauzènes, C. (2022). Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23.

Gentile, C. and Orabona, F. (2012). On multilabel classification and ranking with partial feedback. *Advances in Neural Information Processing Systems*, 25.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):149–170.

Guo, F., Liu, C., and Wang, Y. M. (2009). Efficient multiple-click models in web search. In *Proceedings of the second acm international conference on web search and data mining*, pages 124–131.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192.

Hong, J., Kveton, B., Katariya, S., Zaheer, M., and Ghavamzadeh, M. (2022). Deep hierarchy in bandits. *arXiv preprint arXiv:2202.01454*.

Karatzoglou, A., Baltrunas, L., and Shi, Y. (2013). Learning to rank for recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 493–494.

Kveton, B., Meshi, O., Qin, Z., and Zoghi, M. (2022). On the value of prior in online learning to rank. In *The 25th International Conference on Artificial Intelligence and Statistics*.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015). Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776. PMLR.

Lagrée, P., Vernade, C., and Cappe, O. (2016). Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. (2018). Toprank: A practical algorithm for online stochastic ranking. *Advances in Neural Information Processing Systems*, 31.

Li, C., Kveton, B., Lattimore, T., Markov, I., de Rijke, M., Szepesvári, C., and Zoghi, M. (2020). Bubblerank: Safe online learning to re-rank via implicit click feedback. In *Uncertainty in Artificial Intelligence*, pages 196–206. PMLR.

Liu, T.-Y. et al. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Liu, W., Li, S., and Zhang, S. (2018). Contextual dependent click bandit algorithm for web recommendation. In *International Computing and Combinatorics Conference*, pages 39–50. Springer.

Liu, Y., Devraj, A. M., Van Roy, B., and Xu, K. (2022). Gaussian imagination in bandit learning. *arXiv preprint arXiv:2201.01902*.

Lu, X. and Van Roy, B. (2019). Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Meng, Y., Karimzadehgan, M., Zhuang, H., and Metzler, D. (2020). Separate and attend in personal email search. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 429–437.

Mitra, B. and Craswell, N. (2017). Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. Available at https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.

Neu, G., Olkhovskaya, J., Papini, M., and Schwartz, L. (2022). Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *arXiv preprint arXiv:2205.13924*.

Qin, T. and Liu, T.-Y. (2013). Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*.

Qin, Z., Yan, L., Zhuang, H., Tay, Y., Pasumarthi, R. K., Wang, X., Bendersky, M., and Najork, M. (2021). Are neural rankers still outperformed by gradient boosted decision trees? In *International Conference on Learning Representations*.

Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530.

Santara, A., Aggarwal, G., Li, S., and Gentile, C. (2022). Learning to plan variable length sequences of actions with a cascading bandit click model of user feedback. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 767–797. PMLR.

Shen, J., Karimzadehgan, M., Bendersky, M., Qin, Z., and Metzler, D. (2018). Multi-task learning for email search ranking with auxiliary query clustering. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2127–2135.

Steele, J. M. (2004). *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press.

Tagami, Y., Ono, S., Yamamoto, K., Tsukamoto, K., and Tajima, A. (2013). Ctr prediction for contextual advertising: Learning-to-rank approach. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, pages 1–8.

Vial, D., Sanghavi, S., Shakkottai, S., and Srikant, R. (2022). Minimax regret for cascading bandits. *arXiv preprint arXiv:2203.12577*.

Zamani, H., Bendersky, M., Zhang, M., and Wang, X., editors (2017). *Situational Context for Ranking in Personal Search*.

Zhang, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857.

Zhong, Z., Chueng, W. C., and Tan, V. Y. F. (2021). Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66.

Zoghi, M., Tunys, T., Ghavamzadeh, M., Kveton, B., Szepesvari, C., and Wen, Z. (2017). Online learning to rank in stochastic click models. In *International Conference on Machine Learning*, pages 4199–4208. PMLR.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. (2016). Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*.

A GAUSSIAN TS PROOFS

Theorem 1 (Gaussian TS)**.** *The Bayes regret of Algorithm 1 with prior $\mathcal{N}(0,1)$ satisfies*

$$\mathcal{BR}(T) \leq 22K\sqrt{LT}\ .$$

Proof of Theorem 1. Let

$$U_{k,t} = \hat{\nu}_{k,t} + \sqrt{\frac{\log_+(\frac{T}{LT_{k,t}})}{T_{k,t}}}\ , \tag{4}$$

where $\log_+(x) = \log(x)\mathbf{1}(x \geq 1)$. Now, notice that

$$\begin{aligned} \mathbf{E}_t[U_{A^*(k),t}] &= \mathbf{E}_t[U_{\text{argmax}^k(\nu),t}] \\ &= \mathbf{E}_t[U_{\text{argmax}^k(\hat{\nu}),t}] \\ &= \mathbf{E}_t[U_{A_t(k),t}]\ , \end{aligned} \tag{5}$$

where $\text{argmax}^k(\cdot)$ is the k'th largest element operator. In the first and last equations, we used the definition of U and the second equation is by the construction of the algorithm, as ν and $\tilde{\nu}_t$ are identically distributed conditionally on $\mathcal{H}_t$.

Now, by Kveton et al. (2022, Lemma 7) we know that

$$\begin{aligned} \mathbf{E}_t[\mathcal{R}_t] &= \mathbf{E}_t\Big[\prod_{k\in A^*}(1-\nu_k) - \prod_{k\in A_t}(1-\nu_k)\Big] \\ &= \mathbf{E}_t\Big[\sum_{k=1}^{K}\Big(\prod_{j=1}^{k-1}(1-\nu_{A^*(j)})\Big)(\nu_{A^*(k)} - \nu_{A_t(k)})\Big(\prod_{j=k+1}^{K}(1-\nu_{A_t(j)})\Big)\Big] \\ &\leq \mathbf{E}_t\Big[\sum_{k=1}^{K}\nu_{A^*(k)} - \nu_{A_t(k)}\Big]\ . \end{aligned} \tag{6}$$

Thus, for CM we have

$$\begin{aligned} \mathcal{BR}(T) &:= \mathbf{E}[\sum_{t=1}^{T} r_{A^*} - r_{A_t}] \\ &\leq \mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\nu_{A^*(k)} - \nu_{A_t(k)}\Big] && \text{(by Eq. (6))} \\ &= \mathbf{E}\Big[\sum_{t=1}^{T}\mathbf{E}_t\Big[\sum_{k=1}^{K}\nu_{A^*(k)} - U_{A^*(k),t} + U_{A_t(k),t} - \nu_{A_t(k)}\Big]\Big] && \text{(by Eq. (5) and the tower rule of expectation)} \\ &= \mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{E}_t\Big[\nu_{A^*(k)} - U_{A^*(k),t}\Big] + \sum_{k=1}^{K}\mathbf{E}_t\Big[U_{A_t(k),t} - \nu_{A_t(k)}\Big]\Big] \end{aligned}$$

Let's assume $L \leq T$, and let $\delta_0 = 2\sqrt{\frac{L}{T}}$. We know by the integration by parts formula

$$\mathbf{E}_t\Big[\nu_{A^*(k)} - U_{A^*(k),t}\Big] \leq \delta_0 + \int_{\delta_0}^{1}\mathbf{P}(\nu_i - U_{i,t} \geq u)\mathrm{d}u$$

We can use the following inequality from Audibert and Bubeck (2010) which holds for any $i \in [L]$

$$\mathbf{P}(\nu_i - U_{i,t} \geq u) \leq \frac{16L}{Tu^2}\log(\sqrt{\frac{T}{L}}u) + \frac{1}{Tu^2/L - 1}$$

then using the same reasoning as in Bubeck and Liu (2013, Theorem 1), we can show by integration that

$$\begin{aligned}\int_{\delta_0}^{1} \frac{16L}{Tu^2} \log(\sqrt{\frac{T}{L}}u)\mathrm{d}u &= \left[-\frac{16L}{Tu} \log\left(e\sqrt{\frac{T}{L}}u\right)\right]_{\delta_0}^{1} \\ &\leq \frac{16L}{T\delta_0} \log\left(e\sqrt{\frac{T}{L}}\delta_0\right) \\ &= 8(1+\log 2)\sqrt{\frac{L}{T}}\end{aligned}$$

and

$$\begin{aligned}\int_{\delta_0}^{1} \frac{1}{Tu^2/L-1}\mathrm{d}u &= \left[-\frac{1}{2}\sqrt{\frac{L}{T}} \log\left(\frac{\sqrt{\frac{T}{L}}u+1}{\sqrt{\frac{T}{L}}u-1}\right)\right]_{\delta_0}^{1} \\ &\leq \frac{1}{2}\sqrt{\frac{L}{T}} \log\left(\frac{\sqrt{\frac{T}{L}}\delta_0+1}{\sqrt{\frac{T}{L}}\delta_0-1}\right) \\ &= \frac{\log 3}{2}\sqrt{\frac{L}{T}}\end{aligned}$$

Thus, putting these together we get

$$\mathbf{E}_t\Big[\nu_{A^*(k)} - U_{A^*(k),t}\Big] \leq 13\sqrt{\frac{L}{T}}, \qquad \forall k \in [K]\ .$$

Then taking the Bayes regret expectation over ν, and Lemma 9, we get

$$\mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{E}_t\Big[\nu_{A^*(k)} - U_{A^*(k),t}\Big]\Big] \leq 13K\sqrt{LT}\ . \tag{7}$$

Similarly, by integration by parts, we again know for any $k \in [K]$

$$\sum_{t=1}^{T}\mathbf{E}_t\Big[U_{A_t(k),t} - \nu_{A_t(k)}\Big] \leq \delta_0 T + \int_{\delta_0}^{\infty}\sum_{t=1}^{T}\mathbf{P}(U_{A_t(k),t} - \nu_{A_t(k)} \geq u)\mathrm{d}u\ .$$

By definition of $U_{A_t(k),t}$ and a union bound style argument we can write for any $k \in [K]$

$$\sum_{t=1}^{T}\mathbf{P}(U_{A_t(k),t} - \nu_{A_t(k)} \geq u) \leq \sum_{t=1}^{T}\sum_{i=1}^{L}\mathbf{P}\left(\hat{\nu}_{i,t} + \sqrt{\frac{\log_+(\frac{T}{LT_{i,t}})}{T_{i,t}}} - \nu_i \geq u\right)\ .$$

Now we fix i, let $s(u) = \lceil\frac{3\log(\frac{Tu^2}{L})}{u^2}\rceil$ for $u \geq \delta_0$, and $c = 1 - \frac{1}{\sqrt{3}}$, then we can write

$$\begin{aligned}\sum_{t=1}^{T}\mathbf{P}\left(\hat{\nu}_{i,t} + \sqrt{\frac{\log_+(\frac{T}{LT_{i,t}})}{T_{i,t}}} - \nu_i \geq u\right) &\leq s(u) + \sum_{t=s(u)}^{T}\mathbf{P}(\hat{\nu}_{i,t} - \nu_i \geq cu) \\ &\leq s(u) + \sum_{t=s(u)}^{T}\exp(-2tc^2u^2)\mathbf{1}(u \leq 1/c) \qquad \text{(Hoeffding's inequality)} \\ &\leq s(u) + \frac{\exp(-12c^2\log 2)}{1-\exp(-2c^2u^2)}\mathbf{1}(u \leq 1/c)\ .\end{aligned}$$

Now note that

$$\int_{\delta_0}^{+\infty} \frac{3\log\left(\frac{Tu^2}{L}\right)}{u^2}\mathrm{d}u \leq 3(1+\log(2))\sqrt{\frac{T}{L}} \leq 5.1\sqrt{\frac{T}{L}}$$

and by the fact that $1-\exp(-u) \geq u - u^2/2$ for $u \geq 0$ we can write

$$\begin{aligned}
\int_{\delta_0}^{1/c} \frac{1}{1-\exp(-2c^2u^2)}\mathrm{d}u &= \int_{\delta_0}^{1/(2c)} \frac{1}{1-\exp(-2c^2u^2)}\mathrm{d}u + \int_{1/(2c)}^{1/c} \frac{1}{1-\exp(-2c^2u^2)}\mathrm{d}u \\
&\leq \int_{\delta_0}^{1/(2c)} \frac{1}{2c^2u^2 - 2c^4u^4}\mathrm{d}u + \frac{1}{2c(1-\exp(-1/2))} \\
&\leq \int_{\delta_0}^{1/(2c)} \frac{2}{3c^2u^2}\mathrm{d}u + \frac{1}{2c(1-\exp(-1/2))} \\
&= \frac{2}{3c^2\delta_0} - \frac{4}{3c} + \frac{1}{2c(1-\exp(-1/2))} \\
&\leq 1.9\sqrt{\frac{T}{L}}.
\end{aligned}$$

which altogether means for any $k \in [K]$ we get

$$\sum_{t=1}^{T} \mathbf{E}_t\left[U_{A_t(k),t} - \nu_{A_t(k)}\right] \leq 9\sqrt{LT}\ .$$

Now again by Lemma 9 we get

$$\mathbf{E}\left[\sum_{k=1}^{K}\sum_{t=1}^{T} \mathbf{E}_t\left[U_{A_t(k),t} - \nu_{A_t(k)}\right]\right] \leq 9K\sqrt{LT}\ . \tag{8}$$

Finally, Eq. (7) and Eq. (8) together show $\mathcal{BR}(T) \leq 22K\sqrt{LT}$ for Algorithm 1. □

A.1 Gaussian vs. Beta TS

In this section we compare the Gaussian posterior to the Bernoulli posterior in a TS algorithm. The Gaussian version updates its posterior as given in Eq. (2), while Beta does it as follows:

$$\alpha_{i,t} = \alpha_i + \sum_{\ell=1}^{t} Y_{i,\ell}O_{i,t}, \quad \beta_{i,t} = \beta_i + T_{i,t} - \sum_{\ell=1}^{t} Y_{i,\ell}O_{i,t}\ ,$$

and the mean and variance of Beta posterior are (lose notation here)

$$\hat{\nu}_{i,t} = \frac{\alpha_{i,t}}{\alpha_{i,t}+\beta_{i,t}} \tag{9}$$

$$\hat{\sigma}^2_{i,t} = \frac{\alpha_{i,t}\beta_{i,t}}{(\alpha_{i,t}+\beta_{i,t})^2(\alpha_{i,t}+\beta_{i,t}+1)}\ . \tag{10}$$

It is not hard to see both posterior parameters of item i converge to mean $O(\frac{\sum_{\ell=1}^{t} Y_{i,\ell}O_{i,t}}{T_{i,t}})$ and variance $O(1/T_{i,t})$ for large t, almost with the same rate. However, the difference is in their certainty and concentration around the true mean. Fig. 7 illustrates the posterior distribution for different true means ($\nu \in \{0.1, 0.5, 0.9\}$) after $O_{i,t} = 100$ (so the variance is $1/100$). As we can see for the extreme cases of attractiveness (ν close to 0 or 1), the Beta distribution does not concentrate and its uncertainty is very high (indeed, the entropy = ∞ for Beta posterior). The Gaussian posterior is however suitable for different values of ν. The extreme values of ν are more important in ranking problems as the best items are highly attractive (large ν) and choosing a very unattractive item (small ν) could result in linear regret. Also, in practice usually ν is small, so the posterior must be very accurate around zero. We leave the theoretical establishment of this observation and its benefits in overcoming the prior misspecification to the future work.

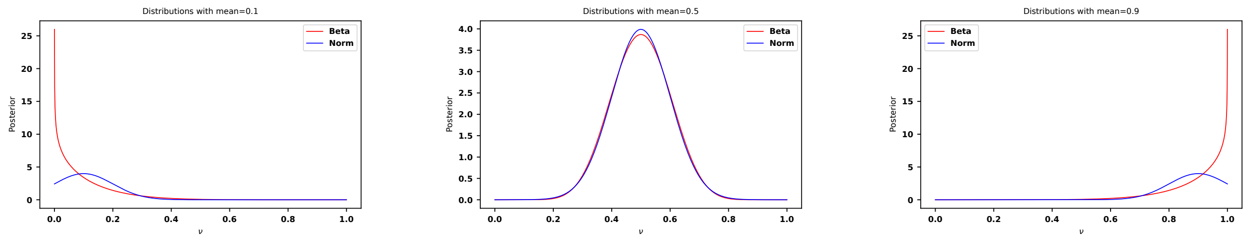

Figure 7: Concentration of Beta vs. Gaussian posterior after 100 samples.

B LINEAR TS PROOFS

Theorem 2 (Algorithm 2 Regret)**.** *Under Assumptions 1 to 3, the Bayes regret of Algorithm 2 is* $\mathcal{BR}(T) = \tilde{O}\left(d^{3/2}K\sqrt{T}\right)$, *where $\tilde{O}$ hides logarithmic terms.*

Proof of Theorem 2. First we write out the conditional step regret as follows

$$\mathbf{E}_t[\mathcal{R}_t] = \mathbf{E}_t\left[\sum_{k=1}^{K} x_{A^*(k),t}^\top \theta - x_{A_t(k),t}^\top \theta\right] \tag{11}$$

Notice that by the construction of the algorithm we know

$$\mathbf{E}_t[x_{A^*(k),t}^\top \theta] = \mathbf{E}_t[x_{A_t(k),t}^\top \tilde{\theta}_t] \ , \tag{12}$$

for all k, t. This is because given $\mathcal{H}_t$, $\tilde{\theta}_t$ is identically distributed as θ and A_t is a deterministic function of $\tilde{\theta}_t$ (See Abeille and Lazaric (2017, Section 4) and Bubeck and Liu (2013, Proof of Theorem 1, Step 1)).

Now, for the cascade model we have

$$\begin{aligned}
\mathcal{BR}(T) &:= \mathbf{E}[\sum_{t=1}^{T} r_{A_t^*} - r_{A_t}] \\
&\leq \mathbf{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} x_{A^*(k),t}^\top \theta - x_{A_t(k),t}^\top \theta\right] && \text{(by Eq. (11))} \\
&= \mathbf{E}\left[\sum_{t=1}^{T} \mathbf{E}_t\left[\sum_{k=1}^{K} x_{A_t(k),t}^\top \tilde{\theta}_t - x_{A_t(k),t}^\top \theta\right]\right] && \text{(by Eq. (12) and tower rule)} \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbf{E}_t\left[x_{A_t(k),t}^\top(\tilde{\theta}_t - \hat{\theta}_t)\right] + \mathbf{E}_t\left[x_{A_t(k),t}^\top(\hat{\theta}_t - \theta)\right]\right] && (13)
\end{aligned}$$

Let's define

$$\begin{aligned}
\mathcal{E}_{\hat{\theta},t} :=& \{\forall x \in \mathcal{X}_t, |x^\top(\hat{\theta}_t - \theta)| \leq \|x\|_{V_t^{-1}}\beta_t(\delta_t)\} \ , \\
\mathcal{E}_{\tilde{\theta},t} :=& \{\forall x \in \mathcal{X}_t, |x^\top(\hat{\theta}_t - \tilde{\theta})| \leq \|x\|_{V_t^{-1}}\gamma_t(\delta_t)\} \ .
\end{aligned}$$

where $\gamma_t(\delta) = \beta_t(\delta)\sqrt{cd\log(c'd/\delta)}$ with c, c' constants such that $\mathbf{P}\left(\|Y_{i,t} - x_{i,t}^\top\theta\| \leq \sqrt{cd\log(c'd/\delta)}\right) \geq 1-\delta$. Note that based on Assumption 1 and Hoeffding's inequality, we can set $c = c' = 2$ (see Abeille and Lazaric (2017, Appendix A, Example 2)).

Abbasi-Yadkori et al. (2011, Theorem 2) shows that $\mathbf{P}(\mathcal{E}_{\hat{\theta},t}) = 1 - \delta_t$, so a union bound along with Lemma 5 shows $\mathbf{P}(\cap_{t=1}^{T}\mathcal{E}_{\hat{\theta},t}) \geq 1 - \delta(\frac{\log T}{2} + 1)$.

(Abeille and Lazaric, 2017, Lemma 1, the proof in the Appendix D) states that $\mathbf{P}(\mathcal{E}_{\tilde{\theta},t}) = 1 - \delta_t$. Again, a union bound along with Lemma 5 shows $\mathbf{P}(\cap_{t=1}^T \mathcal{E}_{\tilde{\theta},t}) \geq 1 - \delta(\frac{\log T}{2} + 1)$.

Thus, using the high probability bounds above, under $\cap_{t=1}^T \mathcal{E}_{\hat{\theta},t}$ and $\cap_{t=1}^T \mathcal{E}_{\tilde{\theta},t}$ we can bound Eq. (13) with

$$\begin{aligned}
&\mathbf{E}\Bigg[\sum_{t=1}^T\sum_{k=1}^K \mathbf{1}(\mathcal{E}_{\tilde{\theta},t})\mathbf{E}_t\Big[x_{A_t(k),t}^\top(\tilde{\theta}_t - \hat{\theta}_t)\Big] + \mathbf{1}(\mathcal{E}_{\hat{\theta},t})\mathbf{E}_t\Big[x_{A_t(k),t}^\top(\hat{\theta}_t - \theta)\Big]\Bigg] \\
&\leq \mathbf{E}\Bigg[\sum_{t=1}^T\sum_{k=1}^K \gamma_t(\delta_t)\|x_{A_t(k),t}\|_{V_t^{-1}} + \beta_t(\delta_t)\|x_{A_t(k),t}\|_{V_t^{-1}}\Bigg] \\
&\leq \mathbf{E}\Bigg[K\sum_{t=1}^T 2\gamma_t(\delta_t) d\log(1+t/\lambda) + 2\beta_t(\delta_t) d\log(1+t/\lambda)\Bigg] && \text{(by Proposition 6)} \\
&\leq 4dK(\gamma_T(\delta_1) + \beta_T(\delta_1))\sum_{t=1}^T \log(1+t/\lambda) && (14) \\
&\leq 4dK\big(\gamma_T(\delta/2) + \beta_T(\delta/2)\big)\sqrt{2Td\log(1+T/\lambda)}\ , && \text{(By Lemma 9)}
\end{aligned}$$

where Eq. (14) is by $\beta_t(\delta)$ and $\gamma_t(\delta)$ being decreasing in δ and increasing in t, and $\delta_1 > \delta_t, \forall t > 1$. Bounding the regret under the complement of the event that the bounds do not hold with probability $\delta(\log T + 2)$, we get

$$\mathcal{BR}(T) \leq \big(1 - \delta(\log T + 2)\big)4dK\big(\gamma_T(\delta/2) + \beta_T(\delta/2)\big)\sqrt{2Td\log(1+T/\lambda)} + \delta(\log T + 2)T\ .$$

Now, take $\delta = \frac{1}{T(\log T + 2)}$, which simplifies the previous inequality to

$$\begin{aligned}
\mathcal{BR}(T) &\leq (1 - 1/T)4dK\big(\gamma_T(\tfrac{1}{2T(\log T+2)}) + \beta_T(\tfrac{1}{2T(\log T+2)})\big)\sqrt{2Td\log(1+T/\lambda)} + 1 \\
&= O\Bigg(dK\sqrt{\log(dT(\log T))}\sqrt{\log\big((1+T/\lambda)^{d/2}T(\log T)\big)}\sqrt{dT\log(1+T/\lambda)}\Bigg) \\
&= \tilde{O}\Big(d^{3/2}K\sqrt{T}\Big)\ .
\end{aligned}$$

□

Lemma 5 (Doubling Trick)**.**

$$\sum_{t=1}^T 1/2^{\max(1,\lceil\log t\rceil)} \leq \frac{\log T}{2} + 1$$

Proof.

$$\begin{aligned}
\sum_{t=1}^T 1/2^{\max(1,\lceil\log t\rceil)} &= 2/2 + 2/4 + 4/8 + \cdots + 2^{\lfloor\log T\rfloor - 1}/2^{\lfloor\log T\rfloor} + (T - 2^{\lfloor\log T\rfloor})/2^{\lceil\log T\rceil} \\
&= 1 + (\lfloor\log T\rfloor - 1)/2 + (T - 2^{\lfloor\log T\rfloor})/2^{\lceil\log T\rceil} \\
&\leq 1 + \lfloor\log T\rfloor/2 - 1/2 + T/2^{\lceil\log T\rceil} - 1/2 \\
&\leq \lfloor\log T\rfloor/2 + T/2^{\log T} \\
&\leq \frac{\log T}{2} + 1
\end{aligned}$$

□

Proposition 6. *Let $\lambda \geq 1$, for any arbitrary sequence $(x_1, x_2, \ldots, x_t) \in (\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_t)$*

$$\sum_{s=1}^t \|x_s\|^2_{V_s^{-1}} \leq 2\log\frac{\det(V_{t+1})}{\det(\lambda I)} \leq 2d\log\Big(1 + \frac{t}{\lambda}\Big). \tag{15}$$

C LOGISTIC TS PROOFS

In this section, we use the results in Filippi et al. (2010); Abeille and Lazaric (2017) to extend the results of the linear case to the GLM case. First, we restate the following result from Appendix F of Abeille and Lazaric (2017). Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, Y_t, \cdots, x_t, Y_t))$ be the sigma algebra generated by the prior knowledge $\mathcal{F}_1$ and the history up to round t.

Proposition 7 (Proposition 11, Abeille and Lazaric (2017))**.** *For any $\delta \in (0,1)$ and $t \geq 1$, under Assumptions 2 to 5, for any $\mathcal{F}_t$-adapted sequence of contexts $(x_1, \cdots, x_t)$, the estimation $\hat{\theta}_t$ from Eq. (3) is such that*

$$\|\hat{\theta}_t - \theta\|_{V_t} \leq \frac{\beta_t(\delta)}{\kappa}$$

and

$$\forall x \in \mathbb{R}^d, \quad \|\mu(x^\top \hat{\theta}) - \mu(x^\top \theta)\| \leq \frac{\ell \beta_t(\delta)}{\kappa} \|x\|_{V_t^{-1}}$$

with probability $1 - \delta$.

Theorem 3 (Algorithm 3 Regret)**.** *The Bayes regret of Algorithm 3 under Assumptions 2 to 4 satisfies*

$$\begin{aligned} \mathcal{BR}(T) \leq K\frac{\ell}{\kappa}\big(\beta_t(\delta') + \gamma_t(\delta')d\big)\sqrt{2Td\log(1 + \frac{T}{\lambda})} \\ + K\frac{2\gamma_t(\delta')}{0.1\kappa}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}} \end{aligned}$$

with probability at least $1 - \delta$ where $\delta' = \frac{\delta}{4T}$. Here $\gamma_t(\delta) = \beta_t(\delta)\sqrt{cd\log(c'd/\delta)}$ with c, c' constants such that $\mathbf{P}\left(\|Y_{i,t} - x_{i,t}^\top\theta\| \leq \sqrt{cd\log(c'd/\delta)}\right) \geq 1 - \delta$. We also show by choosing a proper δ and γ the regret is $\tilde{O}\left(K\frac{\ell}{\kappa}\sigma^2 d\sqrt{Td}\right)$.

Proof of Theorem 3. Starting with Eq. (6), we know

$$\begin{aligned} \mathbf{E}_t[\mathcal{R}_t] &\leq \mathbf{E}_t\left[\sum_{k=1}^K \nu_{A_t^*(k)} - \nu_{A_t(k)}\right] \\ &= \mathbf{E}_t\left[\sum_{k=1}^K \mu(x_{A_t^*(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\theta)\right] \\ &= \mathbf{E}_t\left[\sum_{k=1}^K \mu(x_{A_t^*(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\tilde{\theta}_t) - \sum_{k=1}^K \mu(x_{A_t(k),t}^\top\tilde{\theta}_t) - \mu(x_{A_t(k),t}^\top\theta)\right] \\ &\leq \mathbf{E}_t\left[\sum_{k=1}^K \mu(x_{A_t^*(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\tilde{\theta}_t) - \sum_{k=1}^K \ell\|x_{A_t(k),t}\|_{V_t^{-1}}\|\tilde{\theta}_t - \theta\|_{V_t^{-1}}\right]. \end{aligned} \tag{16}$$

where Eq. (16) is by definition of ℓ and Cauchy-Schwarz inequality. The second term is bounded the same way as in Theorem 1 of Abeille and Lazaric (2017) using Proposition 7 for each $k \in [K]$

$$\sum_{k=1}^K \ell\|x_{A_t(k),t}\|_{V_t^{-1}}\|\tilde{\theta}_t - \theta\|_{V_t^{-1}} \leq K\frac{\ell}{\kappa}\big(\beta_t(\delta') + \gamma_t(\delta')d\big)\sqrt{2Td\log(1 + \frac{T}{\lambda})} \tag{17}$$

For the first term in Eq. (16) we can use the properties of TS and have

$$\mathbf{E}_t\left[\sum_{k=1}^K \mu(x_{A_t^*(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\tilde{\theta}_t)\right] = \mathbf{E}_t\left[\sum_{k=1}^K \mu(x_{A_t(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\tilde{\theta}_t)\right] \;\; \forall k \in [K]$$

Now, note that if $\sup_{x\in\mathcal{X}} x\theta - \sup_{x\in\mathcal{X}} x\tilde{\theta}_t \geq 0$, then by definition of ℓ and removing the absolute value we get

$$\mu(x_{A_t(k),t}^\top\theta) - \mu(x_{A_t(k),t}^\top\tilde{\theta}_t) \leq \ell\big(\sup_{x\in\mathcal{X}} x\theta - \sup_{x\in\mathcal{X}} x\tilde{\theta}_t\big) \;\; \forall k \in [K]$$

and otherwise, by definition of κ we (always) have

$$\mu(x_{A_t(k),t}^\top \theta) - \mu(x_{A_t(k),t}^\top \tilde{\theta}_t) \leq \kappa\big(\sup_{x\in\mathcal{X}} x\theta - \sup_{x\in\mathcal{X}} x\tilde{\theta}_t\big) \;\; \forall k \in [K]$$

By the bound on R^{TS} in Theorem 1 of Abeille and Lazaric (2017) we know

$$\sup_{x\in\mathcal{X}} x\theta - \sup_{x\in\mathcal{X}} x\tilde{\theta}_t \leq \frac{2\gamma_t(\delta')}{0.1\kappa}\mathbf{E}_t\big[\|x_{A_t(k),t}\|_{V_t^{-1}}\big] \;\; \forall k \in [K]$$

(Note that p from Abeille and Lazaric (2017) is replaced by 0.1 accordingly) where $\delta' = \frac{\delta}{4T}$. Next we can use Proposition 2 of Abeille and Lazaric (2017) and get

$$\mathbf{E}_t\Big[\sum_{t=1}^T \sum_{k=1}^K \|x_{A_t(k),t}\|_{V_t^{-1}}\Big] \leq K\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}\;,$$

which means

$$\sum_{t=1}^T \mathbf{E}_t\Big[\sum_{k=1}^K \mu(x_{A_t(k),t}^\top \theta) - \mu(x_{A_t(k),t}^\top \tilde{\theta}_t)\Big] \leq K\frac{2\gamma_t(\delta')}{0.1\kappa}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}} \tag{18}$$

Putting it all together (Eq. (17) and Eq. (18)) we get

$$\mathcal{BR}(T) = \sum_{t=1}^T \mathbf{E}_t[\mathcal{R}_t] \leq K\frac{\ell}{\kappa}\big(\beta_t(\delta') + \gamma_t(\delta')d\big)\sqrt{2Td\log(1+\frac{T}{\lambda})}$$
$$+K\frac{2\gamma_t(\delta')}{0.1\kappa}\sqrt{\frac{8T}{\lambda}\log\frac{4}{\delta}}$$

holds with probability at least $1-\delta$. Now, if we take $\delta = \frac{1}{T}$ and $\gamma = 1$, the regret upper bound is

$$\tilde{O}\bigg(K\frac{\ell}{\kappa}\sigma^2 d\sqrt{Td}\bigg)$$

□

C.1 Removing the Dependence on κ and ℓ

In this section, we prove the result in Theorem 4 for the Sigmoid function as the link function.

Theorem 4 (Algorithm 3 Regret without κ)**.** *If Assumptions 2 to 4 hold and μ is the Sigmoid function, then the Bayes regret of Algorithm 3 satisfies*

$$\mathcal{BR}(T) \leq K\sqrt{2L(d\log(2ST+1)+1)} = \tilde{O}(K\sqrt{dLT})$$

Proof of Theorem 4. We define the *partial lifted information gain* (Neu et al., 2022) as

$$\rho_{t,k} := \frac{(\mathbf{E}_t[r_{A^*(k)}(\theta, x_t) - r_{A_t(k)}(\theta, x_t)])^2}{\mathrm{I}_t(\theta; Y_{t,A_t(k)})}$$

where $r_{A(k)}(\theta, x)$ is the probability of click (mean reward) for item $A(k)$ under context x. Also,

$$\mathrm{I}_t(\theta; Y_{t,A(k)}) := \mathbf{E}_t\big[d_{\text{KL}}(\mathbf{P}(Y_{t,A(k)}|\theta, x_t, A(k), \mathcal{H}_t)\|\mathbf{P}(Y_{t,A(k)}|x_t, A(k), \mathcal{H}_t))\big]$$

is the mutual information between θ and $Y_{t,A(k)}$ conditioned on the history, context x_t, and action A_t. Here $d_{\text{KL}}(p\|q)$ is the KL divergence between p and q distributions.

Since $\mu(\cdot) \in [0,1]$, we can write the step regret conditioned on the history using Kveton et al. (2022, Lemma 7) as follows

$$\begin{aligned}\mathbf{E}_t[\mathcal{R}_t] &\le \mathbf{E}_t\Big[\sum_{k=1}^{K} \mu(x_{A^*(k),t}^\top \theta) - \mu(x_{A_t(k),t}^\top \theta)\Big] \\ &= \mathbf{E}_t\Big[\sum_{k=1}^{K} \sqrt{\rho_{t,k} \mathrm{I}_t(\theta; Y_{t,A(k)})}\Big] , \end{aligned} \tag{19}$$

where in the last equality we used the definition of ρ and r. Now by tower rule, we have

$$\begin{aligned}\mathcal{BR}(T) &= \mathbf{E}\big[\sum_{t=1}^{T} \mathbf{E}_t[\mathcal{R}_t]\big] \\ &\le \sqrt{\mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K} \rho_{t,k}\Big] \mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{I}_t(\theta; Y_{t,A_t(k)})\Big]} .\end{aligned}$$

where the inequality is by Lemma 9 and Eq. (19). Now by Lemma 8 we get

$$\mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K} \rho_{t,k}\Big] \le 2KL .$$

Let Θ_0 denote the set such that $\theta \in \Theta_0$. By Neu et al. (2022, Lemma 6), for any k and $\varepsilon > 0$ we get $\mathbf{E}\Big[\sum_{t=1}^{T} \mathrm{I}_t(\theta; Y_{t,A_t(k)})\Big] \le \log\left(\mathcal{N}_\varepsilon(\Theta_0)\right) + \varepsilon T$, where $\mathcal{N}_\varepsilon(\Theta_0)$ is the ε-covering number of Θ_0 w.r.t ℓ_2-norm. By Assumption 3 and the standard result on the covering number of the Euclidean ball in $\mathbb{R}^d$ we know $\mathcal{N}_\varepsilon(\Theta_0) \le (\frac{2S}{\epsilon} + 1)^d$. Thus, choosing $\varepsilon = 1/T$ we get

$$\mathbf{E}\Big[\sum_{t=1}^{T}\sum_{k=1}^{K} \mathrm{I}_t(\theta; Y_{t,A_t(k)})\Big] \le K(d\log(2ST+1)+1) .$$

Putting all the pieces together, we get the following bound

$$\mathcal{BR}(T) \le K\sqrt{2L(d\log(2ST+1)+1)} = \tilde{O}(K\sqrt{dLT}) .$$

□

Lemma 8 (Bounding the Information Ratio). *If $Y_{i,t} \in \{0,1\}$ for all i, t, then, $\rho_{t,k} \le 2\sum_{i\in\mathcal{L}} \mathbf{E}_t[\hat{r}_t(x_t, i)]$ for all $t \ge 1$ and $k \in \{1, \cdots, K\}$.*

Proof of Lemma 8. Through Fenchel-Young inequality, we prove this using an idea similar to Neu et al. (2022, Lemma 5). First, for $p, q \in [0,1]$ let's remind the following definition of d_{KL} for the Bernoulli distribution

$$d_{\mathrm{KL}}(p||q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} ,$$

where $0\log 0 = 0$ convention is used. Also, let $\hat{r}_{A_t(k),t}(x) := \mathbf{E}_t[r_{A_t(k)}(\theta, x)]$ be the posterior mean reward at round t for item at position k of the action, for parameter context tuple (θ, x). The Legendre–Fenchel conjugate of d_{KL} with respect to its first argument is defined for all $u \in \mathbb{R}$ as

$$d^*_{\mathrm{KL}}(u\|q) := \sup_{p\in[0,1]} (pu - d_{\mathrm{KL}}(p\|q)) = \log(1 + q(e^u - 1)) , \tag{20}$$

where the second equality and the inequality follow from Proposition 10. Let $\pi_{t,k}(i|x_t) = \mathbf{P}(A_t(k) = i|x_t, \mathcal{H}_t)$ be the probability of recommending item $i \in \mathcal{L}$ at position k at round t. Given $\mathcal{H}_t$ at round t, we let $\overset{\mathrm{d}}{=}_t$ denote equality in distribution. Now define the pseudo-regret at step t for position k in the action as $\mathcal{R}_{t,k} = r_{A^*(k)}(\theta, x_t) - r_{A_t(k)}(\theta, x_t)$, then for any $\eta > 0$ and $k \in \{1, \cdots, K\}$

$$
\begin{aligned}
\mathbf{E}_t[\mathcal{R}_{t,k}] &= \mathbf{E}_t\left[r_{A^*(k)}(\theta, x_t) - r_{A_t(k)}(\theta, x_t)\right] \\
&= \mathbf{E}_t\left[r_{A_t(k)}(\tilde{\theta}_t, x_t) - r_{A_t(k)}(\theta, x_t)\right] && (\text{As } (\theta, A_t^*) \overset{\mathrm{d}}{=}_t (\tilde{\theta}_t, A_t)) \\
&= \mathbf{E}_t\left[r_{A_t(k)}(\tilde{\theta}_t, x_t) - \hat{r}_{A_t(k),t}(x_t)\right] && (\text{By conditional independence of } \theta \text{ and } A_t) \\
&= \mathbf{E}_t\left[\left(\sum_{i\in\mathcal{L}} \mathbf{1}(A_t(k) = i)\frac{\eta\pi_{t,k}(i|x_t)}{\eta\pi_{t,k}(i|x_t)} r_i(\tilde{\theta}_t, x_t)\right) - \hat{r}_{A_t(k),t}(x_t)\right] \\
&\le \mathbf{E}_t\Bigg[\Bigg(\eta\sum_{i\in\mathcal{L}} \pi_{t,k}(i|x_t)\Bigg(d_{\mathrm{KL}}(r_i(\tilde{\theta}_t, x_t)\|\hat{r}_{i,t}(x_t)) \\
&\qquad + d^*_{\mathrm{KL}}\left(\frac{-\mathbf{1}(A_t(k) = i)}{\eta\pi_{t,k}(i|x_t)}\middle\|\hat{r}_{i,t}(x_t)\right)\Bigg)\Bigg) - \hat{r}_{A_t(k),t}(x_t)\Bigg] \\
& && (\text{By Eq. (20) where } u \leftarrow \tfrac{-\mathbf{1}(A_t(k)=i)}{\eta\pi_{t,k}(i|x_t)},\ p \leftarrow r_i(\tilde{\theta}_t, x_t), \text{ and } q \leftarrow \hat{r}_{i,t}(x_t)) \\
&\le \mathbf{E}_t\Bigg[\Bigg(\eta\sum_{i\in\mathcal{L}} \pi_{t,k}(i|x_t)\Bigg(d_{\mathrm{KL}}(r_i(\tilde{\theta}_t, x_t)\|\hat{r}_{i,t}(x_t)) - \frac{\mathbf{1}(A_t(k) = i)}{\eta\pi_{t,k}(i|x_t)}\hat{r}_{i,t}(x_t) \\
&\qquad + \frac{\mathbf{1}(A_t(k) = i)}{2\eta^2\pi_{t,k}(i|x_t)^2}\hat{r}_{i,t}(x_t)\Bigg)\Bigg) - \hat{r}_{A_t(k),t}(x_t)\Bigg] && (\text{By Proposition 10}) \\
&= \mathbf{E}_t\left[\sum_{i\in\mathcal{L}} \pi_{t,k}(i|x_t) d_{\mathrm{KL}}(r_i(\theta, x_t)\|\hat{r}_{i,t}(x_t)) + \frac{1}{2\eta}\hat{r}_t(x_t, i)\right] \\
& && (\text{By tower rule and } \mathbf{E}_t[\mathbf{1}(A_t(k) = i)] = \pi_{t,k}(i|x_t) \text{ and } \theta \overset{\mathrm{d}}{=}_t \tilde{\theta}_t) \\
&= \eta \mathrm{I}_t(\theta; Y_{A_t(k),t}) + \frac{1}{2\eta}\sum_{i\in\mathcal{L}} \mathbf{E}_t[\hat{r}_t(x_t, i)]
\end{aligned}
$$

Minimizing over η we get

$$
\mathbf{E}_t[\mathcal{R}_t] \le \sqrt{2\mathrm{I}_t(\theta; Y_{A_t(k),t})\sum_{i\in\mathcal{L}} \mathbf{E}_t[\hat{r}_t(x_t, i)]}\ , \tag{21}
$$

which readily gives the result. □

D Alternative Logistic TS Algorithm

In this section, we develop a logistic LTR algorithm that does not require solving equations like Eq. (3). Therefore, this algorithm is more computationally attractive. We define $\bar{O} = 1$ if $Y_{i,t} = 1$ and $\bar{O} = -1$ if $Y_{i,t} = 0$. Algorithm 4 lays out the algorithm with step size α. Employing standard analysis of online Newton step algorithms like an adaptation of Hazan et al. (2007); Gentile and Orabona (2012) we can derive similar guarantees for our algorithm as Algorithm 3. A close example of this is the algorithm in Santara et al. (2022).

E TECHNICAL TOOLS

Lemma 9 (Cauchy-Schwartz Inequality in Euclidean Vector Space)**.** *For all vectors u and v in $\mathbb{R}^d$ we know*

$$
\sum_{i=1}^{d} u_i v_i \le \sqrt{\sum_{i=1}^{d} u_i^2}\sqrt{\sum_{i=1}^{d} v_i^2}\ .
$$

This is a classical inequality that has several proofs (Steele, 2004).

Proposition 10 (Neu et al. (2022), Proposition 1.)**.** *For any $u \le 0$ and $q \in [0, 1]$:*

$$
d^*_{KL}(u\|q) \le q\left(u + \frac{u^2}{2}\right).
$$

Algorithm 4 Logistic TS for LTR using Newton Steps

Input: Step size α
Initialize: $c_1 = 1$, $V_1 = KI$, and $\hat{\theta}_1 = 0$
for $t = 1, \ldots, T$ **do**
  Receive the current context x_t.
  Sample $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_{c_t}, V_{c_t}^{-1})$ # Posterior sample
  $A_t \in \arg\max_{\substack{a \subset \mathcal{L} \\ |a|=K}} \sum_{i \in a} x_{i,t}^\top \tilde{\theta}_t$ # Recommend the top posterior samples
  Observe Y_t and gather $\mathcal{H}_{t+1}$ and $(\mathcal{T}_{i,t})_{i \in \mathcal{L}}$.
  for $k \leq k_t$ **do**
    $V_{c_t+k-1} \leftarrow V_{c_t+k-2} + O_{k,t} x_{k,t} x_{k,t}^\top$
    $\hat{\theta}_{c_t+k} \leftarrow \hat{\theta}_{c_t+k-1} + \alpha\mu\left(-\bar{O}_{k,t}\theta_{c_t+k-1}^\top x_{i,t}\right)\bar{O}_{k,t} V_{c_t+k-1}^{-1} x_{k,t}$
  end for
  $c_{t+1} = c_t + k_t$
end for

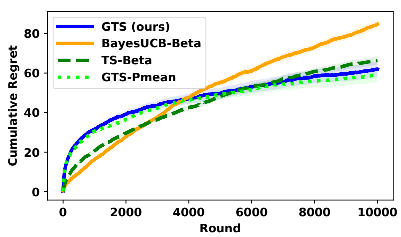

Figure 8: Synthetic non-contextual setting with a subset of algorithms to highlight the differences.

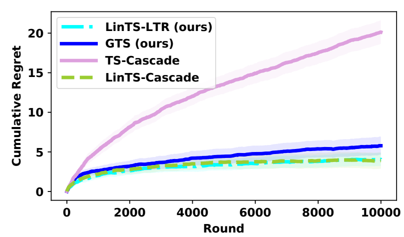

Figure 9: Linear model experiment same as Fig. 3 but some of the algorithms were removed to focus on the difference between the high performing ones.

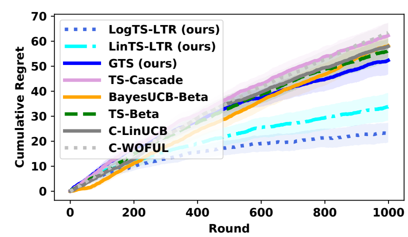

Figure 10: The logistic model experiment showing the first 1000 steps.

Proof.

$$d^*_{\mathrm{KL}}(u\|q) = \log(1 + q(e^u - 1)) \leq q(e^u - 1) \leq q\left(u + \frac{u^2}{2}\right),$$

where the first inequality is from $\log(1 + x) \leq x$ for any $x > -1$, and the second inequality is from $e^x \leq 1 + x + \frac{x^2}{2}$ for any $x \leq 0$. □

F FURTHER EXPERIMENTS AND DETAILS

Experimental setup configuration: For the synthetic experiments, we used a machine with AMD EPYC 7B12, x86-64 processor with 48 cores and 200 GB memory, and each one took around 5 minutes to complete.

Dataset Experiment Details: For Web30K and Istella, we remove the features which have a standard deviation less than 10^{-6} after normalizing across the datasets.

Further experiments: Fig. 11 shows a truncated version of Fig. 5 for a closer look into the early rounds of the experiment. As we can see, `LogTS-LTR` outperforms the other algorithms.

Fig. 13 shows that `GTS` has a competitive performance compared to `TS`.

Figs. 15 and 16 demonstrate the prior initialization where `BetaMean` is the mean of the (true) Beta prior, and `BetaVar` is its variance. The legend denotes the [prior mean, variance] for the Gaussian prior of `GTS-P`. We can observe that the prior does not have a huge impact on our `GTS-P` algorithm (another confirmation to its robustness w.r.t prior misspecification). It seems, however, the correct prior ([`BetaMean, BetaVar`]) shows slightly better performance in the early stages.

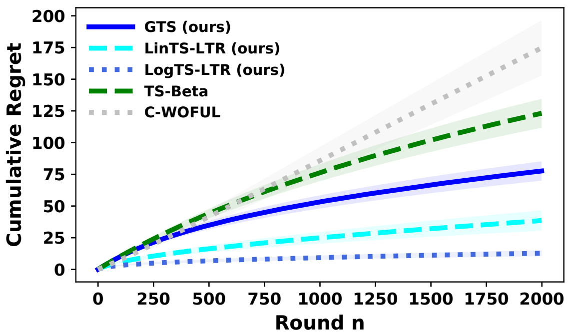

Figure 11: Web30k experiment truncated at 2000 steps.

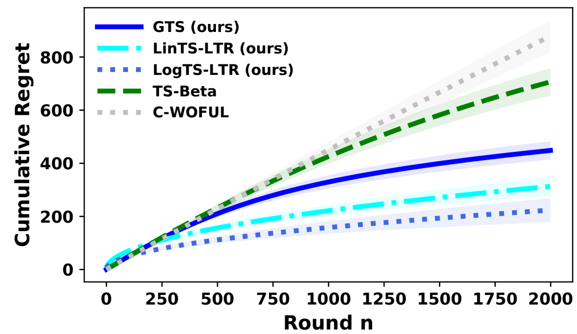

Figure 12: Istella experiment truncated at 2000 steps.

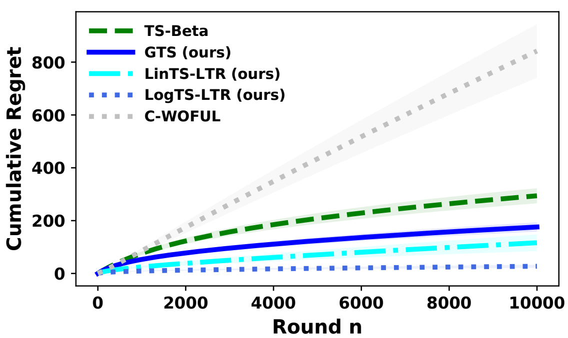

Figure 13: The Web30k experiment with all the algorithms.

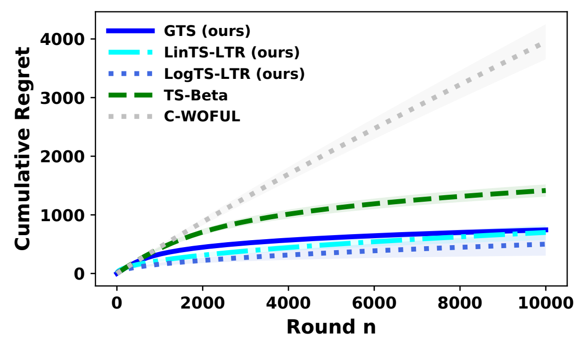

Figure 14: The Istella experiment with all the algorithms.

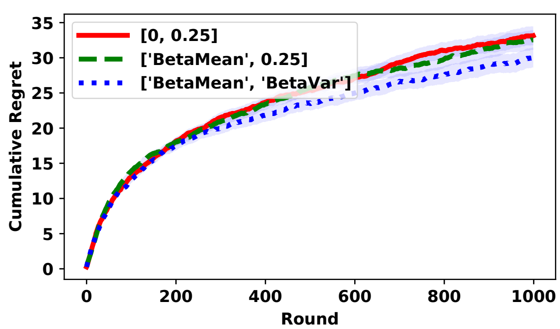

Figure 15: The effect of prior initialization on `GTS-P` in a Beta environment, $T = 1000$.

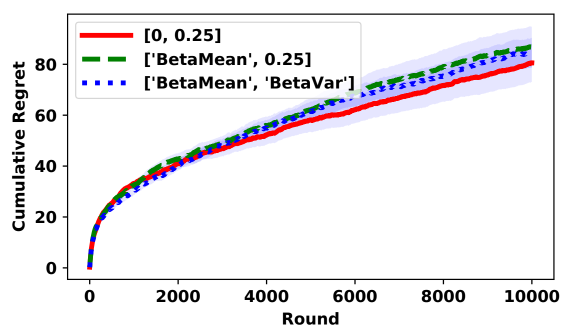

Figure 16: The effect of prior initialization on `GTS-P` in a Beta environment, $T = 10000$.