Thompson Sampling for Constrained Bandits

Anonymous authors
Paper under double-blind review

Keywords: Bandits with Knapsacks, Thompson Sampling, Conservative Bandits.

Summary

Contextual bandits model sequential decision-making where an agent balances exploration and exploitation to maximize long term cumulative rewards. Many real-world applications, such as online advertising and inventory pricing, impose additional resource constraints while in high-stakes settings like healthcare and finance, early-stage exploration can pose significant risks. The Contextual Bandits with Knapsacks (CBwK) framework extends contextual bandits to incorporate resource constraints while the Contextual Conservative Bandit (CCB) framework ensures that performance remains above $(1+\alpha)$ times the performance of a predefined safe baseline. Although Upper Confidence Bound (UCB) based methods exist for both setups, a Thompson Sampling (TS) based approach has not been explored. This gap in the literature motivates the need to study TS for constrained settings, further reinforced by the fact that Thompson sampling often demonstrates superior empirical performance in the unconstrained setting. In this work we consider linear CBwK and CCB setups and design Thompson sampling algorithms `LinCBwK-TS` and `LinCCB-TS` respectively. We provide a $\tilde{\mathcal{O}}\big((\frac{\text{OPT}}{B}+1)m\sqrt{T}\big)$ regret for `LinCBwK-TS` where OPT is the optimal value and B is the total budget. Further, we show that `LinCCB-TS` has a regret bounded by $\tilde{\mathcal{O}}\big(\sqrt{T}\min\{m^{3/2}, m\sqrt{\log K}\} + \Delta_h/\alpha r_l(\Delta_l + \alpha r_l)\big)$ and maintains the performance guarantee with high probability where Δ_h and Δ_l are the upper and lower bounds on the baseline gap and r_l is a lower bound on baseline reward.

Contribution(s)

1. We provide a Thompson Sampling Algorithm for Linear Contextual Bandits with Knapsacks and prove a high probability regret bound.
   Context: Previous work looked at an Upper Confidence Bound (UCB) approach.
2. We provide a Thompson Sampling Algorithm for Linear Contextual Conservative Bandits and prove a high probability regret bound along with showing that it satisfies a performance constraint.
   Context: Previous work looked at an Upper Confidence Bound (UCB) approach.

Thompson Sampling for Constrained Bandits

Anonymous authors
Paper under double-blind review

Abstract

Contextual bandits model sequential decision-making where an agent balances exploration and exploitation to maximize long-term cumulative rewards. Many real-world applications, such as online advertising and inventory pricing, impose additional resource constraints, while in high-stakes settings like healthcare and finance, early-stage exploration can pose significant risks. The Contextual Bandit with Knapsack (CBwK) framework extends contextual bandits to incorporate resource constraints while the Contextual Conservative Bandit (CCB) framework ensures that performance of the learner remains above $(1-\alpha)$ times the performance of a predefined safe baseline. Although Upper Confidence Bound (UCB) based methods exist for both setups, a Thompson Sampling (TS) based approach has not been explored. This gap in the literature motivates the need to study TS for constrained settings, further reinforced by the fact that TS often demonstrates superior empirical performance in the unconstrained setting. In this work, we consider linear CBwK and CCB settings and design TS algorithms `LinCBwK-TS` and `LinCCB-TS` respectively. We provide a $\tilde{\mathcal{O}}\big((\frac{\text{OPT}}{B}+1)m\sqrt{T}\big)$ regret for `LinCBwK-TS` where OPT is the optimal value and B is the total budget. Further, we show that `LinCCB-TS` has a regret bounded by $\tilde{\mathcal{O}}\big(\sqrt{T}\min\{m^{3/2}, m\sqrt{\log K}\} + m^3\Delta_h/\alpha r_l(\Delta_l+\alpha r_l)\big)$ and maintains the performance guarantee with high probability, where Δ_h and Δ_l are the upper and lower bounds on the baseline gap and r_l is a lower-bound on the baseline reward.

1 Introduction

Contextual bandit is a fundamental model in sequential decision-making wherein an agent must balance the exploration of unknown actions with the exploitation of actions believed to be optimal (Langford & Zhang, 2007; Lattimore & Szepesvári, 2020; Slivkins, 2022). At every round, the learner observes K separate context vectors, each corresponding to a possible action (arm). Based on the collected history, the learner then selects an action and receives a reward signal corrupted by noise. The learner's goal is to maximize the total reward accumulated over T rounds, or equivalently, to minimize the regret when compared to the best action selection strategy in hindsight. In many real-world applications, there are additional *constraints* on how actions can be selected while interacting with the environment. For instance, the learner may have to manage limited resources or ensure that performance does not fall below an existing baseline.

The *Bandit with Knapsack (BwK)* framework (Badanidiyuru et al., 2013; Agrawal & Devanur, 2016) incorporates mechanisms to handle resource limitations within the contextual bandit setting. When an action is selected, the learner observes a reward along with a consumption vector, and the objective is to maximize the cumulative rewards while ensuring that the cumulative consumptions are below a given budget. This problem arises in various real-world applications. For example, in clinical trials, researchers must balance the exploration of new treatments while adhering to constraints such as the availability of medical facilities, drugs, and patient participation. In online advertising, ad placements are not only influenced by user engagement but also by advertisers' budgets, which limit the number of times an ad can be displayed. Similarly, retailers conducting price experimentation must manage both consumer demand and inventory constraints to maximize revenue.

Additionally, complex AI systems, such as assistive robots and autonomous vehicles, require substantial computational resources, and their training is often limited by available processing power and energy constraints.

Further in many real-world applications, exploring new strategies can often lead to substantial risks, especially in high-stakes domains such as healthcare, finance, and online advertising. While standard bandit algorithms eventually converge to an optimal policy, their early performance can be unpredictable and unsafe, making them impractical for deployment in many settings. To address this, *conservative bandits* impose safety constraints that require the algorithm's cumulative rewards to remain within a controlled range of an existing baseline policy (Wu et al., 2016; Kazerouni et al., 2017). The learner is then required to find a good policy while ensuring that cumulative reward at every round is greater than the cumulative reward of the baseline by a $(1-\alpha)$ factor.

Thompson Sampling (TS) has demonstrated strong empirical performance across a wide range of sequential decision-making problems. It was not studied in great depth for several decades after its introduction by Thompson (1933) until Chapelle & Li (2011) and Scott (2010) successfully showed that it matches the state-of-the-art results, and in many cases significantly outperforms alternative algorithms. Subsequently TS was applied to a wide variety of domains including website optimization (Hill et al., 2017), recommendation systems (Kawale et al., 2015), and revenue management (Ferreira et al., 2018). Also see Russo et al. (2020) for a detailed tutorial on TS. In recent times, TS has also been combined with neural networks to provide improved performance (Riquelme et al., 2018; Zhang et al., 2021b). Although Upper Confidence Bound (UCB) based methods have been developed for both Contextual Bandit with Knapsack (CBwK) (Agrawal & Devanur, 2016) and Contextual Conservative Bandit (CCB) frameworks (Kazerouni et al., 2017; Garcelon et al., 2020), a Thompson Sampling (TS) based approach remains unexplored. Further, given that TS often demonstrates superior empirical performance in the unconstrained setting, it is natural to investigate whether it can be effectively extended to constrained decision-making problems. This motivates our study, where we design and analyze TS-based algorithms for CBwK and CCB frameworks.

We outline our main contributions below.

1. **Linear CBwK:** We consider the linear CBwK problem and design a TS-based primal-dual algorithm `LinCBwK-TS` (see Algorithm 1). We prove that `LinCBwK-TS` enjoys a regret bound of $\mathcal{O}\Big(\big(\frac{\text{OPT}}{B}+1\big)m\sqrt{T\log(dT/\delta)\log(TK)}\Big)$ with high probability (see Theorem 3.1) where OPT is the optimal value and B is the total budget (see Section 3 for details).
2. **Linear CCB:** We consider the linear CCB problem and design a TS-based algorithm with a safety condition, `LinCCB-TS` (see Algorithm 2). Subsequently we prove that `LinCCB-TS` simultaneously satisfies the performance guarantee in high probability with respect to the existing baseline (cf. Theorem 4.2) and enjoys a regret bound of $\tilde{\mathcal{O}}\big(\sqrt{T}\min\{m^{3/2}, m\sqrt{\log K}\}+m^3\Delta_h/\alpha r_l(\Delta_l+\alpha r_l)\big)$.

2 Related Work

Contextual Bandits. Research on bandit algorithms, particularly in the contextual setting, has progressed in several directions. Early work on linear bandits focused on developing exploration strategies with theoretical regret bounds. Abe et al. (2003), Chu et al. (2011), and Abbasi-Yadkori et al. (2011) explored methods based on linear models, leading to algorithms that performed well in different settings. Agrawal & Goyal (2012) analyzed regret guarantees for Thompson Sampling in the multi-armed case, later extending these results to the linear setting with formal guarantees (Agrawal & Goyal, 2013). Following these developments, researchers examined extensions to non-linear models. Generalized linear bandits (GLBs) were introduced by Filippi et al. (2010) and Li et al. (2017), incorporating non-linearity through a link function while retaining a linear dependence on contextual information.

The use of deep learning in contextual bandits has also been explored. Some approaches relied on deep neural networks (DNNs) for feature extraction, with a linear model trained on top of the last hidden layer of the network (Lu & Van Roy, 2017; Zahavy & Mannor, 2020; Riquelme et al., 2018). While these methods showed promising empirical performance, they lacked theoretical regret guarantees. To address this, Zhou et al. (2020) introduced NeuralUCB, which used neural networks along with UCB-based exploration and provided regret bounds. Zhang et al. (2021a) extended this approach to Thompson Sampling, incorporating ideas from neural tangent kernels (NTKs) (Jacot et al., 2018; Allen-Zhu et al., 2019) and the effective dimension $\tilde{d}$.

Foster & Rakhlin (2020) introduced SquareCB, which connects contextual bandit regret to online regression with square loss. Foster & Krishnamurthy (2021) later proposed FastCB, which modifies SquareCB using KL loss to achieve a data-dependent regret bound. Additionally, Simchi-Levi & Xu (2020) showed that in the stochastic setting, an offline regression oracle can achieve optimal regret with significantly fewer calls than online regression-based approaches. A recent work by Deb et al. (2024a) extended both SquareCB and FastCB using neural networks and also demonstrated regret bounds by Zhou et al. (2020) and Zhang et al. (2021a) are $\Omega(T)$ in the worst case, even against an oblivious adversary.

Constrained Bandits. Bandits with constraints require an agent to optimize rewards while adhering to operational limits, and come in several variants. Bandits with Knapsacks (BwK) was introduced by (Badanidiyuru et al., 2013) for the multi-armed bandit (MAB) setting. In BwK, pulling an arm yields both a reward and a consumption, with the goal of maximizing cumulative reward before depleting a limited budget. This formulation was later extended to the linear contextual bandits (Agrawal & Devanur, 2016), concave rewards and convex constraints (Agrawal & Devanur, 2014a), adversarial bandits (Immorlica et al., 2022b; Sivakumar et al., 2022) and dueling bandits (Deb et al., 2024b). For general reward and consumption functions, (Slivkins et al., 2023; Han et al., 2023) provided sub-linear regret bounds using inverse gap weighting techniques (Abe & Long, 1999; Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021). The online optimization with knapsacks problem is another closely related problem, where feedback is available for all actions after a decision is made. This problem has been studied through online linear and convex programming techniques (Agrawal & Devanur, 2014b; Mahdavi et al., 2012).

Another related setting is conservative bandits, where the agent must ensure that its reward at each round does not fall below a predefined fraction of a baseline policy and was introduced in (Wu et al., 2016). Existing methods primarily rely on Upper Confidence Bound (UCB)-based approaches (Wu et al., 2016; Kazerouni et al., 2017; Garcelon et al., 2020) and a recent paper studied the inverse gap weighted version Deb et al. (2025). A different constrained bandit model involves stage-wise constraints, where the expected reward of an action must not exceed a given threshold at each round. Unlike BwK, which enforces constraints cumulatively, this formulation imposes per-step limitations. (Amani et al., 2019; Moradipari et al., 2019) studied this setting for linear bandits, proposing explore-exploit and Thompson Sampling-based algorithms, respectively.

3 Contextual Bandits with Knapsacks

There are K actions labeled by $[K] = \{1, \ldots, K\}$ and a budget $B \in \mathbb{R}_+$. In each round t, a context vector $\mathbf{x}_t(a) \in [0,1]^m$ is observed for each action $a \in [K]$, and the learner chooses an action a_t (or a "no-op" option). Subsequently, a reward $r_t(a_t) \in [0,1]$ and a d-dimensional consumption vector $\mathbf{v}_t(a_t) \in [0,1]^d$ are observed, both drawn independently from the past history $\mathcal{F}_t$. The d elements of the vector $\mathbf{v}_t(a_t)$ are the consumptions associated with d different types of resources.

Assumption 3.1. *There exist unknown parameters $\mu_* \in [0,1]^m$ and $W_* \in [0,1]^{m \times d}$ such that for each action a, the conditional expectation of the reward and consumption are given by*

$$\mathbb{E}[r_t(a) \mid \mathbf{x}_t(a), \mathcal{F}_{t-1}] = \mu_*^\top \mathbf{x}_t(a), \qquad \mathbb{E}[\mathbf{v}_t(a) \mid \mathbf{x}_t(a), H_{t-1}] = W_*^\top \mathbf{x}_t(a) \; .$$

The objective of the agent is to design a policy that maximizes the T-step total reward $\sum_{t=1}^{T} r_t(a_t)$ subject to the following budget constraint: *the cumulative consumption must not exceed the budget*

Algorithm 1 : `LinCBwK-TS` (**Lin**ear **C**ontextual **B**andits **w**ith **K**napsacks - **T**hompson **S**ampling)

1: **Initialize** θ_1 according to the OCO algorithm and Z such that $\frac{\text{OPT}}{B} \leq Z \leq O\left(\frac{\text{OPT}}{B}+1\right)$
2: **for** $t = 1, \ldots, T$ **do**
3: Observe $\mathbf{x}_t(a), \forall a \in [K]$, and compute the parameter estimates according to (1)
4: Sample $\tilde{\mu}(t) \sim \mathcal{N}\left(\hat{\mu}(t), v_t^2 \hat{\Sigma}(t)\right)$ and $\tilde{w}(t) \sim \mathcal{N}\left(\hat{W}(t)^\top \theta_t, v_t^2 \hat{\Sigma}(t)\right)$
5: **Play arm**

$$a_t := \operatorname*{argmax}_{a \in [K]} \left\{ \mathbf{x}_t(a)^\top \left(\tilde{\mu}(t) - Z\, \tilde{w}(t)\right) \right\}$$

6: Observe $r_t(a_t)$ and $\mathbf{v}_t(a_t)$.
7: **If** there exists a $j \in [d]$ such that $\sum_{t'=1}^{t} \mathbf{v}_{t'}(a_{t'}) \cdot e_j \geq B$, **then exit**
8: Use $\mathbf{x}_t(a_t)$, $r_t(a_t)$, and $\mathbf{v}_t(a_t)$ to obtain $\hat{\mu}(t+1)$ and $\hat{W}(t+1)$ using (1).
9: **Update** θ_{t+1} via the OCO algorithm with

$$g_t(\theta_t) := \theta_t \cdot \left(\mathbf{v}_t(a_t) - \tfrac{B}{T}\mathbb{1}\right)$$

10: **end for**

any dimension, i.e., $\sum_{t=1}^{T} \mathbf{v}_t(a_t) \leq B\mathbb{1}$. Consider a policy π that is *context dependent but non-adaptive*. When the realized context is $X \in \mathcal{X}$, the policy π assigns a probability distribution $\pi(X) \in \Delta^{K+1}$ over the K arms plus a "no-op." We define the *expected reward* and the *expected consumption* of π as

$$r(\pi) := \mathbb{E}_{X\sim\mathcal{D}}\left[\mu_*^\top X\, \pi(X)\right], \qquad v(\pi) := \mathbb{E}_{X\sim\mathcal{D}}\left[W_*^\top X\, \pi(X)\right].$$

and the *optimal static policy* as $\pi^* := \arg\max_\pi \; T\, r(\pi)$ subject to $T\, v(\pi) \leq B\,\mathbb{1}$.

Regret: Suppose OPT $= Tr(\pi^*)$. We define the regret of an algorithm that plays the actions $\{a_t\}_{t\in[T]}$ as

$$\texttt{Reg}_{\texttt{BwK}}(T) := \text{OPT} - \sum_{t=1}^{\tau} r_t(a_t).$$

3.1 Algorithm

We will use a primal-dual approach as in Agrawal & Devanur (2016); Sivakumar et al. (2022); Immorlica et al. (2022a). After observing the rewards and consumption vectors until time t, the algorithm constructs a least-squares estimate of the reward parameter μ_* and each row of the consumption parameter matrix W_* as follows:

$$\hat{\mu}(t) = \hat{\Sigma}(t)^{-1}\left(\sum_{\tau=1}^{t-1} \mathbf{x}_\tau(a_\tau) r(a_\tau)\right), \qquad \hat{W}(t)_j = \hat{\Sigma}(t)^{-1}\left(\sum_{\tau=1}^{t-1} \mathbf{x}_\tau(a_\tau)\mathbf{v}(a_\tau)_j\right), \;\; \forall j \in [d]$$

$$\text{where} \qquad \hat{\Sigma}(t) = I + \sum_{\tau=1}^{t-1} \mathbf{x}_\tau(a_\tau)\mathbf{x}_\tau(a_\tau)^\top. \tag{1}$$

Our algorithm `LinCBwK-TS` summarized in Algorithm 1 proceeds as follows. In each round t, it begins by initializing the dual variable θ_t using an Online Convex Optimization (OCO) algorithm and sets the scaling factor Z. Upon observing the context vectors $\mathbf{x}_t(a), \forall a \in [K]$, it computes estimates for reward and consumption using Bayesian linear regression. Specifically, it samples $\tilde{\mu}(t) \sim \mathcal{N}\left(\hat{\mu}(t), v_t^2\hat{\Sigma}(t)\right)$ and $\tilde{w}(t) \sim \mathcal{N}\left(\hat{W}(t)^\top\theta_t, v_t^2\hat{\Sigma}(t)\right)$. Note that for the consumption, we do not generate d different Gaussians corresponding to the d different columns of W_*. Instead, we use the current dual variable θ_t and generate one single Gaussian with mean $\hat{W}(t)^\top\theta_t$. This is computationally more efficient and as we will show in the proof, it allows us to derive a regret bound that does not scale exponentially with the number of consumptions d.

Using these estimates, the algorithm selects an action a_t that maximizes the adjusted reward function, incorporating the estimated reward and scaled consumption (line 5). After playing the selected arm, the algorithm observes the actual reward and consumption values (line 6). If at any point the accumulated resource consumption exceeds the allocated budget along any dimension, the algorithm terminates early (line 7). Otherwise, the observed values are used to update the reward and consumption estimates (line 8). The online convex optimization (OCO) algorithm is then advanced by updating θ_t. The OCO algorithm chooses a sequence $(\theta_t)_{t\in[T]}$ that minimizes the OCO regret on $g_t(\theta_t) : \Omega \to \mathbb{R}$ defined as

$$\mathcal{R}(T) := \max_{\theta\in\Omega} \sum_{t=1}^{T} g_t(\theta) - \sum_{t=1}^{T} g_t(\theta_t),$$

where $\Omega = \{\theta : \theta \geq 0, \|\theta\|_1 \leq 1\}$. Subsequently we will refer to $\mathcal{R}(T)$ as the dual regret since it measures the regret on the dual variable θ_t. We make the following assumption on the dual regret. Note that several OCO algorithms, e.g., Online Mirror Descent (OMD), satisfy this assumption (see Hazan 2021).

Assumption 3.2 (Dual Regret). *Suppose $g_t(\theta_t) := \theta_t \cdot \left(\mathbf{v}_t(a_t) - \frac{B}{T}\mathbb{1}\right)$. Then we assume that the sequence $(\theta_t)_{t\in[T]}$ in line 9 of Algorithm 1 has a dual regret that satisfies $\mathcal{R}(T) \leq \sqrt{T\log d}$.*

3.2 Regret Bound for LinCBwK-TS

In the next Theorem we provide a regret upper bound for our algorithm LinCBwK-TS and thereafter provide a proof sketch. For the purposes of clarity the proof of the intermediate lemmas have been pushed to Appendix A.

Theorem 3.1: Regret of LinCBwK-TS (Algorithm 1)

Suppose the rewards and consumptions satisfy Assumption 3.1. Then LinCBwK-TS (Algorithm 1) achieves the following regret bound with probability at least $1-\delta$,

$$\mathtt{Reg}_{\mathtt{BwK}}(T) \leq \mathcal{O}\left(\left(\frac{OPT}{B}+1\right)m\sqrt{T\log(dT/\delta)\log(TK)}\right).$$

Remark 3.1. *The regret bound in Theorem 3.1 matches the regret bound of the UCB based algorithm for the same setting in Agrawal & Devanur (2016) upto logarithmic factors in K.*

Proof Sketch. We provide a proof sketch and refer the readers to Appendix A for a detailed proof.

We start by defining the Lagrangian at time t as $\ell_t(a) = \mathbf{x}_t(a)^\top\mu_* - Z\theta_t^\top\left(\frac{B}{T}\mathbb{1} - W_*^\top\mathbf{x}_t(a)\right)$. Recall that μ_* and W_* are the true reward and consumption parameters, and θ_t is the dual variable at time t in Algorithm 1. Let $\tau \leq T$ be the time-step when Algorithm 1 stops and $(a_t)_{t\in[\tau]}$ be the sequence of actions selected. We consider the following term which measures the difference between the sum of the Lagrangians (up to the stopping time τ) for the optimal policy π^* and the actions $(a_t)_{t\in[\tau]}$, using the dual variables $(\theta_t)_{t\in[\tau]}$ from Algorithm 1 as the Lagrange multipliers:

$$\mathcal{L}_\tau\left((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\right) = \sum_{t=1}^{\tau}\sum_{a\in[K]}\pi^\star(a)\ell_t(a) - \sum_{t=1}^{\tau}\ell_t(a_t).$$

Our proof will proceed in three steps. In the first step we upper bound the above term.

1. **Upper Bound on** $\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big)$.

We start by observing that

$$\begin{aligned}\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big) &= \sum_{t=1}^{\tau}\sum_{a\in[K]}\pi^\star(a)\ell_t(a) - \sum_{t=1}^{\tau}\ell_t(a_t)\\ &\overset{(a)}{\leq} \sum_{t=1}^{\tau}\sum_{a\in[K]}\pi^\star(a)\ell_t(a_t^\star) - \sum_{t=1}^{\tau}\ell_t(a_t) \overset{(b)}{\leq} \sum_{t=1}^{\tau}\ell_t(a_t^\star) - \sum_{t=1}^{\tau}\ell_t(a_t),\end{aligned} \tag{2}$$

where in (a) and (b) we used $a_t^* = \text{argmax}_{a\in[K]}\,\ell_t(a)$ and $\sum_{a\in[K]}\pi^\star(a) = 1$, respectively. The next lemma bounds the conditional expectation of the above term for any $t \in [T]$. We define the following good events:

$$\begin{aligned}E_1^\mu &= \left\{\forall i\in[K], |\mathbf{x}_t(i)^\top\hat{\mu}(t) - \mathbf{x}_t(i)^\top\mu_*| \leq \left(\sqrt{m\ln\left(\frac{t^3}{\delta}\right)}+1\right)\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\right\}\\ E_2^\mu &= \Bigg\{\forall i\in[K], |\mathbf{x}_t(i)^\top\hat{\mu}(t) - \mathbf{x}_t(i)^\top\tilde{\mu}(t)|\\ &\qquad\qquad \leq v_t^2\min\{\sqrt{4m\log t}, \sqrt{4\log(tK)}\}\sqrt{\mathbf{x}_t(i)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(i)}\Bigg\}\end{aligned}$$

Lemma 3.1. *Let Assumption 3.1 holds and $\mathcal{F}_t$ be the history up to time t. Further suppose E_1^μ and E_2^μ hold. Then for all $t > 0$ and $\delta \in (0,1)$, we have*

$$\begin{aligned}&\mathbb{E}\Big[\ell_t(a_t^*) - \ell_t(a_t)\big|\mathcal{F}_{t-1}\Big]\\ &\qquad \leq C\left(\min\{\sqrt{4m\log(t)}, \sqrt{4\log(tK)}\}v_t + \ell_t\right)\left(\mathbb{E}\left[\sqrt{\mathbf{x}_t(a_t)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(a_t)}\right] + \frac{1}{t^2}\right),\end{aligned}$$

where $\ell_t = \sqrt{m\log(t^3/\delta)} + 1$ and $v_t = \sqrt{m\log(t/\delta)}$.

Using Lemma 3.1 and the proof of Theorem 1 in Agrawal & Goyal (2013), we can conclude that with probability at least $1-\delta$, the following inequality holds:

$$\sum_{t=1}^{\tau}\ell_t(a_t^\star) - \sum_{t=1}^{\tau}\ell_t(a_t) \leq \mathcal{O}\left(m\sqrt{T}\left(\min\left\{\sqrt{m}, \sqrt{\log Kd}\right\}\right)\sqrt{\log(T)\log(1/\delta)}\right).$$

Combining with (2) we have the following high probability (w.p. at leat $1-\delta$) upper-bound on $\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big)$:

$$\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big) \leq \mathcal{O}\left(m\sqrt{T}\left(\min\left\{\sqrt{m}, \sqrt{\log Kd}\right\}\right)\sqrt{\log(T)\log(1/\delta)}\right). \tag{3}$$

2. **Lower bound on** $\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big)$: We now lower-bound the Lagrangian difference. To do so, we separately consider the two terms for a fixed time-step t and write

$$\mathcal{L}_\tau\Big((\theta_t)_{t\in[\tau]}, (a_t)_{t\in[\tau]}\Big) = \sum_{t=1}^{\tau}\underbrace{\sum_{a\in[K]}\pi^\star(a)\ell_t(a)}_{I} - \sum_{t=1}^{\tau}\underbrace{\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t\left(\frac{B}{T}\mathbb{1} - W_*^\top\mathbf{x}_t(a_t)\right)}_{II}.$$

Our objective is to relate term I to the optimal value OPT and term II to the observed cost vectors $\mathbf{v}(a_t)$. The latter would be subsequently bounded via the dual regret in step 3. This is true since the OCO update in line 9 of Algorithm 1 uses the observed consumption vectors $\mathbf{v}(a_t)$. We show that both these terms can be bounded by constructing appropriate martingale sequences and using Azuma Hoeffding. The following lemma formalizes this claim.

Lemma 3.2. *Suppose Assumption 3.1 holds. Then, with probability at least* $1-\delta$, *we have*

$$(i) \quad \sum_{t=1}^{\tau}\sum_{a\in[K]}\pi^*(a)\ell_t(a) \geq \tau\frac{OPT}{T} - (Z+2)\sqrt{T\log\frac{2}{\delta}}\,,$$

$$(ii) \quad \sum_{t=1}^{\tau}\ell_t(a_t) \leq \sum_{t=1}^{\tau}\left[\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t^\top\left(\frac{B}{T}\mathbb{1}-\mathbf{v}_t(a_t)\right)\right] + 2Z\sqrt{T\log(1/\delta)}\,.$$

Combining Lemma 3.2 (i) and (ii) we obtain the following high probability lower-bound on the Lagrangian difference:

$$\mathcal{L}_\tau\left((\theta_t)_{t\in[\tau]},(a_t)_{t\in[\tau]}\right) \geq \tau\frac{\text{OPT}}{T} - \sum_{t=1}^{\tau}\left[\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t^\top\left(\frac{B}{T}\mathbb{1}-\mathbf{v}_t(a_t)\right)\right] - (Z+2)\sqrt{T\log\frac{2}{\delta}} - 2Z\sqrt{T\log(1/\delta)}\,. \tag{4}$$

3. **Bounding Final Regret via Dual Regret:** Combining the upper and lower bounds on $\mathcal{L}_\tau\left((\theta_t)_{t\in[\tau]},(a_t)_{t\in[\tau]}\right)$ from Steps 1 and 2 (i.e., (3) and (4)), we may write

$$\tau\frac{\text{OPT}}{T} - \sum_{t=1}^{\tau}\left[\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t^\top\left(\frac{B}{T}\mathbb{1}-\mathbf{v}_t(a_t)\right)\right] \leq (Z+2)\sqrt{T\log\frac{2}{\delta}} + 2Z\sqrt{T\log(1/\delta)} + \mathcal{O}\left(m\sqrt{T}\left(\min\left\{\sqrt{m},\sqrt{\log Kd}\right\}\right)\sqrt{\log(T)\log(1/\delta)}\right).$$

The final step requires us to bound $Z\theta_t^\top\left(\frac{B}{T}\mathbb{1}-\mathbf{v}_t(a_t)\right)$ via the dual regret. Using Lemma 9 in Agrawal & Devanur (2016), we have

$$\theta_t^\top\left(\mathbf{v}_t(a_t)-\frac{B}{T}\mathbb{1}\right) \geq B - \frac{\tau B}{T} - \sqrt{T\log d}\,.$$

Combining with the previous bound, we may write

$$\tau\frac{\text{OPT}}{T} - \sum_{t=1}^{\tau}\mathbf{x}_t(a_t)^\top\mu_* + \left(ZB - Z\frac{\tau B}{T}\right) \leq (Z+2)\sqrt{T\log\frac{2}{\delta}} + 2Z\sqrt{T\log(1/\delta)} + \mathcal{O}\left(m\sqrt{T}\left(\min\left\{\sqrt{m},\sqrt{\log Kd}\right\}\right)\sqrt{\log(dT)\log(1/\delta)}\right).$$

Using the fact that $Z \geq \frac{\text{OPT}}{B}$, the following holds with probability at least $1-\delta$:

$$\text{OPT} - \sum_{t=1}^{\tau}\mathbf{x}_t(a_t)^\top\mu_* \leq \mathcal{O}\left(\left(\frac{\text{OPT}}{B}+1\right)m\sqrt{T\log(dKT/\delta)\log(T)}\right).$$

□

4 Contextual Conservative Bandits

We consider a contextual bandit problem where a learner makes sequential decisions over T time-steps. At each round $t \in [T]$, the learner observes a context vector $\mathbf{x}_t(a) \in [0,1]^m$ for each $a \in [K]$. The learner selects an arm $a_t \in [K]$ and observes the corresponding reward $r_t(a_t) \in [0,1]$. We make the following assumption on the rewards.

Assumption 4.1. *There exists an unknown parameter $\mu_* \in [0,1]^m$ such that for each action a, the conditional expectation of the reward is given by*

$$r_t^*(a) = \mathbb{E}[r_t(a) \mid \mathbf{x}_t(a), \mathcal{F}_{t-1}] = \mu_*^\top \mathbf{x}_t(a) \ .$$

Definition 4.1 (Regret). *The objective of the learner is to minimize the regret, defined as*

$$\texttt{Reg}_{\texttt{CCB}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_t(a_t^*) - r_t(a_t)\right] = \sum_{t=1}^{T} \mu_*^\top \mathbf{x}_t(a_t^*) - \mu_*^\top \mathbf{x}_t(a_t) \ , \tag{5}$$

where $a_t^* = \underset{a\in[K]}{\operatorname{argmax}}\ \mu_*^\top \mathbf{x}_t(a)$ is the optimal arm that maximizes the expected reward in round t.

We assume the presence of a baseline policy π_b, which selects an action $b_t \in [K]$ at each round t and obtains an expected reward of $\mu_*^\top \mathbf{x}_t(b_t)$. This baseline policy represents the default or status quo strategy used by the company, which is known to have a reasonable performance. While the company aims to improve upon this policy, it seeks to limit excessive costs during the optimization process. To enforce this, we introduce the following performance constraint:

Definition 4.2 (Performance Constraint). *At every round t, the cumulative reward of the learner's policy should not be below the $(1-\alpha)$-fraction of the cumulative reward of the baseline policy for some $\alpha > 0$, i.e.,*

$$\sum_{i=1}^{t} \mu_*^\top \mathbf{x}_i(a_i) \geq (1-\alpha) \sum_{i=1}^{t} \mu_*^\top \mathbf{x}_i(b_i) \ , \quad \forall t \in \{1, \ldots, T\} \ . \tag{6}$$

We assume that the expected rewards associated with actions taken by the baseline policy are known. This assumption is reasonable, since this is the default policy of the company and can be further relaxed to the unknown baselines case using a similar analysis as in Kazerouni et al. (2017). Further we make the following assumption on the baseline rewards following Kazerouni et al. (2017); Garcelon et al. (2020).

Assumption 4.2 (Baseline Gap and Bounds). *Let $\Delta_{t,b_t} := \mu_*^\top \mathbf{x}_t(a_t^*) - \mu_*^\top \mathbf{x}_t(b_t)$ represent the baseline gap at time $t \in [T]$. We assume there exist constants $0 \leq \Delta_l \leq \Delta_h$ and $0 < r_l < r_h$ such that for all $t \in [T]$, we have $\Delta_l \leq \Delta_{t,b_t} \leq \Delta_h$ and $r_l \leq r_{t,b_t} \leq y = r_h$.*

4.1 Algorithm

We represent by $\mathcal{S}_t \subseteq [T]$ the subset of time steps up to round t when the Thompson sampling actions were chosen, while $\mathcal{S}_t^c \subseteq [T]$ corresponds to the time steps when the baseline actions were chosen. Further, the sizes of these sets are given by $n_t = |\mathcal{S}_t|$ and $n_t^c = |\mathcal{S}_t^c|$, respectively. Our algorithm `LinCCB-TS` is summarized in Algorithm 2 and proceeds as follows. At each time step t, the learner receives the contexts $\mathbf{x}_t(a)$ for every $a \in [K]$, and computes the parameter estimate for reward using Bayesian linear regression. Specifically, it samples $\tilde{\mu}(t) \sim \mathcal{N}(\hat{\mu}(t), v_t^2 \hat{\Sigma}(t))$ (line 5). Using these estimates, the algorithm selects an action $\tilde{a}_t$ that maximizes the reward function (line 6). However, before committing to the selected action, it verifies a *safety condition* (see line 7) by

Algorithm 2 : `LinCCB-TS` (**Lin**ear **C**ontextual **C**onservative **B**andits - **T**hompson **S**ampling)

1: **Input:** Performance parameter $\alpha > 0, \hat{\mu}(1) = 0, \hat{\Sigma}(1) = \mathbb{I}_{d\times d}, v_t = \sqrt{9m\ln(t/\delta)}$.
2: **Initialize:** $S_0 = \emptyset$
3: **for** $t = 1, 2, 3, \ldots$ **do**
4: Receive contexts $\{\mathbf{x}_t(a)\}_{a\in[K]}$
5: Sample $\tilde{\mu}(t) \sim \mathcal{N}\big(\hat{\mu}(t), v_t\hat{\Sigma}(t)\big)$
6: Compute $\tilde{a}_t = \text{argmax}_{a\in[K]}\, \tilde{\mu}(t)^\top \mathbf{x}_t(a_t)$
7: **if** the *safety condition* in (7) is satisfied **then**
8: Play $a_t = \tilde{a}_t$ and observe reward $r_t(a_t)$.
9: Update $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{t\}$, $\mathcal{S}_t^c = \mathcal{S}_{t-1}^c$, and $\hat{\mu}(t)$ using (8).
10: **else**
11: Play the baseline action $a_t = b_t$ and observe reward $r_t(b_t)$.
12: Update $\mathcal{S}_t = \mathcal{S}_{t-1}$, $\mathcal{S}_t^c = \mathcal{S}_{t-1}^c \cup \{t\}$.
13: **end if**
14: **end for**

checking if the following inequality holds:

$$\sum_{\tau\in\mathcal{S}_{t-1}} \tilde{\mu}(\tau)^\top \mathbf{x}_\tau(a_\tau) + \tilde{\mu}(t)^\top \mathbf{x}_t(\tilde{a}_t) + \sum_{\tau\in\mathcal{S}_{t-1}^c} r_\tau^*(b_\tau)$$
$$- \sum_{t\in S_{\tau-1}} \left(\min\left\{\sqrt{2m\ln\frac{t}{\delta}}, \sqrt{2\ln\frac{tK}{\delta}}\right\} + \sqrt{m\ln\frac{t^3}{\delta}} + 1\right) v_t\sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)}$$
$$\geq (1-\alpha)\sum_{\tau=1}^{t} r_\tau^*(b_\tau) \tag{7}$$

If the condition is satisfied, the selected action is played and the corresponding reward is observed (line 8). Subsequently we update $\mathcal{S}_t$ and $\mathcal{S}_t^c$ and compute the least-squares estimate of the reward parameter μ_* as follows:

$$\hat{\mu}(t) = \hat{\Sigma}(t)^{-1}\left(\sum_{\tau\in\mathcal{S}_t} \mathbf{x}_\tau(a_\tau) r_\tau(a_\tau)\right), \qquad \hat{\Sigma}(t) = \mathbb{I}_{d\times d} + \sum_{\tau\in\mathcal{S}_t} \mathbf{x}_\tau(a_\tau)\mathbf{x}_\tau(a_\tau)^\top. \tag{8}$$

If the *safety condition* in (7) is not satisfied then the baseline action b_t is played, the corresponding reward $r_t(b_t)$ is observed and the sets $\mathcal{S}_t$ and $\mathcal{S}_t^c$ are updated (line 11 and 12).

4.2 Regret Bound for `LinCCB-TS`

Theorem 4.1: Regret of `LinCCB-TS` (Algorithm 2)

Suppose the rewards satisfy Assumption 4.1 and suppose the baseline rewards satisfy Assumption 3.2 holds. With probability at least $1-\delta$, `LinCCB-TS` *(see Algorithm 2) satisfies the performance constraint in eq. (6) and has the following regret bound:*

$$\texttt{Reg}_{\texttt{CCB}}(T) = \mathcal{O}\Bigg(\underbrace{m\sqrt{T}\big(\min\{\sqrt{m}, \sqrt{\log(K)}\}\big)\big(\ln(T) + \sqrt{\ln(T)\ln(4/\delta)}\big)}_{I} + \underbrace{\frac{C\Delta_h}{\alpha r_l(\alpha r_l + \Delta_l)}}_{II}\Bigg)$$

$$\textit{where} \quad C = \mathcal{O}\left(\frac{m^2\ \min\{m, \log K\}\left(\log^2 T + \log T\log(1/\delta)\right)}{\alpha r_l(\Delta_l + \alpha y_l)}\right).$$

Remark 4.1. *Term I has an additional $\sqrt{m}$ dependence when compared to its linear UCB counterpart in Kazerouni et al. (2017). However this is not an artifact of the conservative analysis and is inherited from the Thompson sampling analysis in the unconstrained setting. Similarly, term II has an additional m dependence when compared with term II in Kazerouni et al. (2017). Additionally, there is a $\log^2 T$ dependence, and this appears because the extra buffer (the third term in the lhs) in the safety condition in (7) has a t dependence for the Thompson sampling version.*

Proof Sketch. We provide a proof sketch and refer the readers to Appendix B for a detailed proof.

1. **Regret Decomposition:** Following Kazerouni et al. (2017) we decompose the regret in (5) into two parts using the following lemma.

   Lemma 4.1. *Let Assumptions 4.1 and 4.2 hold. Then, the regret in (5) can be bounded as*

$$\texttt{Reg}(T) \leq \sum_{t \in \mathcal{S}_T} \Big(h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \Big) + n_T^c \Delta_h, \tag{9}$$

   where the set $\mathcal{S}_T$ consists of the rounds until the horizon T when `LinCCB-TS` *played the TS action and $n_T^c = |\mathcal{S}_T^c|$ is the number of times until T when a baseline action was played.*

2. **Upper Bound on n_T^c :** Next, we bound the number of times `LinCCB-TS` played the baseline action. This is determined by our choice of the *safety condition* in (7). We use $n_t := |\mathcal{S}_t|$ and $\tau := \max\{1 \leq t \leq T : a_t = b_t\}$, i.e., the last time step at which `LinCCB-TS` played a baseline action. In the next Lemma we bound n_T^c.

   Lemma 4.2. *Suppose Assumption 4.1 and 4.2 holds. Then, with probability $1 - \delta/4$ the number of times the baseline action is played by* `C-SquareCB` *is bounded as*

$$n_T^c \leq \mathcal{O}\left(\frac{m^2 \ \min\{m, \log K\} \Big(\log^2 T + \log T \log(1/\delta) \Big)}{\alpha r_l (\Delta_l + \alpha y_l)} \right). \tag{10}$$

3. **Final Regret Bound:** The first term in (9) can be bounded using the TS analysis, Theorem 1 from (Agrawal & Goyal, 2013) giving the following lemma.

   Lemma 4.3. *Suppose Assumptions 4.1 holds. Then, for any $\delta > 0$, with with probability $1 - \delta$,* `LinCCB-TS` *guarantees*

$$\sum_{t \in \mathcal{S}_T} h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \leq \mathcal{O}\Big(d\sqrt{n_T} \big(\min\{\sqrt{m}, \sqrt{\log(K)}\} \big) \big(\ln(T) + \sqrt{\ln(T) \ln(4/\delta)} \big) \Big) \tag{11}$$

   Using $n_T \leq T$ and combining (9), (10), (11), and taking a union bound over the high probability events proves the regret bound in Theorem 4.1 holds with probability $1 - \delta$.

4. **Performance Constraint:** Finally we show that the *performance constraint* in (6) is satisfied using the *safety condition* in Line 7 of `LinCCB-TS` .

   Lemma 4.4. *Let Assumptions 4.1 and 4.2 hold. Then, for any $\delta > 0$, with probability $1 - \delta$,* `LinCCB-TS` *satisfies the performance constraint in (6).*

   Taking a union bound over all the high probability events, `LinCCB-TS` simultaneously satisfies the performance constraint in (6) and the regret upper-bound in (11), which concludes the proof.

□

5 Conclusion

In this work, we introduced Thompson Sampling-based algorithms for two constrained contextual bandit settings that have long been missing in the literature: Contextual Bandit with Knapsack (CBwK) and Contextual Conservative Bandit (CCB). We designed `LinCBwK-TS`, a primal-dual TS algorithm for CBwK, and established a high-probability regret bound of $\tilde{\mathcal{O}}\big((\frac{\mathrm{OPT}}{B}+1)m\sqrt{T}\big)$. Similarly, we developed `LinCCB-TS`, a TS-based approach for CCB, proving that it satisfies the required safety constraints with high probability while achieving a regret bound of $\tilde{\mathcal{O}}\big(\sqrt{T}\min m^{3/2}, m\sqrt{\log K} + \Delta_h/\alpha r_l(\Delta_l + \alpha r_l)\big)$. Our results bridge the gap in the literature by demonstrating that TS can be effectively applied to constrained bandit problems, offering an alternative to UCB-based methods. Future work could explore extensions to nonlinear settings, adversarial constraints, and adaptive exploration strategies to further enhance the practical applicability of TS in constrained decision-making scenarios.

Extending these TS algorithms to the more general reward (and consumption) setting using the recently proposed Feel good Thompson Sampling (Zhang, 2021) is left for future work. Combining these with modern deep networks along the lines of (Zhang et al., 2021b) is also left for future work.

Broader Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11. Citeseer, 1999.

Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, pp. 989–1006. ACM, 2014a. DOI: 10.1145/2600057.2602888.

Shipra Agrawal and Nikhil R. Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1405–1424. SIAM, 2014b.

Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3458–3467, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/agrawal12.html.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. *Linear stochastic bandits under safety constraints*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, oct 2013. DOI: 10.1109/focs.2013.30. URL https://doi.org/10.1109%2Ffocs.2013.30.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=5ep85sakT3.

Rohan Deb, Aadirupa Saha, and Arindam Banerjee. Think before you duel: Understanding complexities of preference learning under constrained resources. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4546–4554. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/deb24a.html.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SThJXvucjQ.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Oper. Res.*, 66(6):1586–1602, November 2018. ISSN 0030-364X. DOI: 10.1287/opre.2018.1755. URL https://doi.org/10.1287/opre.2018.1755.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34, 2021.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3962–3969, Apr. 2020. DOI: 10.1609/aaai.v34i04.5812. URL https://ojs.aaai.org/index.php/AAAI/article/view/5812.

Yuxuan Han, Jialin Zeng, Yang Wang, Yang Xiang, and Jiheng Zhang. Optimal contextual bandits with knapsacks under realizability via regression oracles. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp.

5011–5035. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/han23b.html.

Elad Hazan. Introduction to online convex optimization, 2021.

Daniel N. Hill, Houssam Nassif, Yi Liu, Anand Iyer, and S.V.N. Vishwanathan. An efficient bandit algorithm for realtime multivariate optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1813–1821, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. DOI: 10.1145/3097983.3098184. URL https://doi.org/10.1145/3097983.3098184.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022a. ISSN 0004-5411. DOI: 10.1145/3557045. URL https://doi.org/10.1145/3557045.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022b. ISSN 0004-5411. DOI: 10.1145/3557045. URL https://doi.org/10.1145/3557045.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient thompson sampling for online matrix-factorization recommendation. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. NIPS'17, pp. 3913–3922, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3260–3268, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: Online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1): 2503–2528, 2012.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling. *ArXiv*, abs/1911.02156, 2019. URL https://api.semanticscholar.org/CorpusID:207794176.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling, 2020. URL https://arxiv.org/abs/1707.02038.

Steven L. Scott. A modern bayesian look at the multi-armed bandit. *Appl. Stoch. Model. Bus. Ind.*, 26(6):639–658, November 2010. ISSN 1524-1904. DOI: 10.1002/asmb.874. URL https://doi.org/10.1002/asmb.874.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *ArXiv*, abs/2003.12699, 2020.

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20253–20277. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sivakumar22a.html.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2022.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression, 2023.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvari. Conservative bandits. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/wu16.html.

Tom Zahavy and Shie Mannor. Neural linear bandits: Overcoming catastrophic forgetting through likelihood matching, 2020. URL https://openreview.net/forum?id=r1gzdhEKvH.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning, 2021. URL https://arxiv.org/abs/2110.00871.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021a.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021b.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Regret Bound for LinCBwK-TS

Lemma 3.1. *Let Assumption 3.1 holds and $\mathcal{F}_t$ be the history up to time t. Further suppose E_1^μ and E_2^μ hold. Then for all $t > 0$ and $\delta \in (0, 1)$, we have*

$$\mathbb{E}\Big[\ell_t(a_t^*) - \ell_t(a_t)\Big|\mathcal{F}_{t-1}\Big] \\ \leq C\left(\min\{\sqrt{4m\log(t)}, \sqrt{4\log(tK)}\}v_t + \ell_t\right)\left(\mathbb{E}\left[\sqrt{\mathbf{x}_t(a_t)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(a_t)}\right] + \frac{1}{t^2}\right),$$

where $\ell_t = \sqrt{m\log(t^3/\delta)} + 1$ and $v_t = \sqrt{m\log(t/\delta)}$.

Proof of Lemma 3.1. Before we proceed to the proof of Lemma 3.1, we present a lemma that bounds probability of the following good events.

$$E_1^\mu = \left\{\forall i \in [K], |\mathbf{x}_t(i)^\top\hat{\mu}(t) - \mathbf{x}_t(i)^\top\mu_*| \leq \left(\sqrt{m\ln\left(\frac{t^3}{\delta}\right)} + 1\right)\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\right\}$$

$$E_2^\mu = \Big\{\forall i \in [K], |\mathbf{x}_t(i)^\top\hat{\mu}(t) - \mathbf{x}_t(i)^\top\tilde{\mu}(t)| \\ \leq v_t^2\min\{\sqrt{4m\log t}, \sqrt{4\log(tK)}\}\sqrt{\mathbf{x}_t(i)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(i)}\Big\}$$

$$E_1^W = \left\{\forall i \in [K], |\mathbf{x}_i(t)^\top\hat{W}(t)^\top\theta_t - \mathbf{x}_i(t)^\top W_*\theta_t| \leq \left(\sqrt{m\ln\left(\frac{t^3d}{\delta}\right)} + 1\right)\sqrt{\mathbf{x}_t(i)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(i)}\right\}$$

$$E_2^W = \{\forall i \in [K], |\mathbf{x}_i(t)^\top\hat{W}(t)^\top\theta_t - \mathbf{x}_i(t)^\top\tilde{w}(t)| \\ \leq v_t^2\min\{\sqrt{4m\log dt}, \sqrt{4\log(tK)}\}\sqrt{\mathbf{x}_t(i)^\top\hat{\Sigma}(t)^{-1}\mathbf{x}_t(i)}\Big\}$$

Lemma A.1. *For all t, $0 < \delta < 1$ and any filtration $\mathcal{F}_{t-1}$ we have*

$$P(E_1^\mu) \geq 1 - \frac{\delta}{t^2}, \quad P(E_2^\mu|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{t^2}$$
$$P(E_1^W) \geq 1 - \frac{\delta}{t^2}, \quad P(E_2^W|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{t^2}$$

Proof of Lemma A.1. The claims $P(E_1^\mu) \geq 1 - \frac{\delta}{t^2}$ and $P(E_2^\mu|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{t^2}$ follows from Lemma 1 in (Agrawal & Goyal, 2013).

Next, we define $\eta_t = Z\theta_t^\top\left(\hat{W}(t)^\top\mathbf{x}_t(a) - W_*^\top\mathbf{x}_t(a)\right)$, where recall that $\hat{W}(t)_j = \Sigma(t)^{-1}\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a_\tau)\mathbf{v}_\tau(a_\tau)_j$. Now note that

$$Z\theta_t\left(\hat{W}(t) - W_*\right)^\top\mathbf{x}_t(a) \leq Z\|\theta_t\|_1\|(\hat{W}(t) - W_*)^\top\mathbf{x}_t(a)\|_\infty \\ \leq Z\max_{j\in[d]}\left|\left(\hat{W}(t)_j - W_{*j}\right)^\top\mathbf{x}_t(a)\right|$$

Fix $j \in [d]$ and consider the inner product

$$\begin{aligned}\left(\hat{W}(t)_j - W_{*j}\right)^\top \mathbf{x}_t(a) &= \left(\Sigma(t)^{-1}\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a_\tau)\mathbf{v}_\tau(a_\tau)_j - W_{*j}\right)^\top \mathbf{x}_t(a) \\ &= \left(\Sigma(t)^{-1}\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a_\tau)\mathbf{v}_\tau(a_\tau)_j - W_{*j}\right)^\top \mathbf{x}_t(a) + \left(\Sigma(t)^{-1}\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a_\tau)\eta_{\tau,j}\right)^\top \mathbf{x}_t(a) \\ &\le \left(\left\|\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a)\eta_{\tau,j}\right\|_{\Sigma(t)^{-1}} + 1\right)\|\mathbf{x}_t(a)\|_{\Sigma(t)^{-1}}\end{aligned}$$

Now using Theorem 1 of (Abbasi-Yadkori et al., 2011) we have with probability $1-\delta$

$$\left\|\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a)\eta_{\tau,j}\right\|_{\Sigma(t)^{-1}} \le 2\sqrt{m\ln\frac{t}{\delta}}$$

Taking a union bound over all $j \in [d], a \in [K]$, we have with probability $1-\frac{\delta}{t^2}$ for all $a \in [K]$ and $j \in [d]$:

$$\left\|\sum_{\tau=1}^{t-1}\mathbf{x}_\tau(a)\eta_{\tau,j}\right\|_{\Sigma(t)^{-1}} \le \sqrt{m\ln\left(\frac{t^2dK}{\delta}\right)} + 1$$

Therefore, with probability $1-\frac{\delta}{t^2}$ for all $a \in [K], j \in [d]$,

$$\begin{aligned}\left(\hat{W}(t)_j - W_{*j}\right)^\top \mathbf{x}_t(a) &\le \|\mathbf{x}_t(a)\|_{\Sigma(t)^{-1}}\left(\sqrt{m\ln\left(\frac{t^3dK}{\delta}\right)} + 1\right) \\ &= \left(\sqrt{m\ln\left(\frac{t^3dK}{\delta}\right)} + 1\right)\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\end{aligned}$$

Next observe that

$$\begin{aligned}\left|\tilde{w}(t)^\top\mathbf{x}_t(a) - \theta_t^\top\hat{W}(t)^\top\mathbf{x}_t(a)\right| &= \left|\tilde{w}(t)^\top\mathbf{x}_t(a) - \left(\hat{W}(t)\theta_t\right)^\top\mathbf{x}_t(a)\right| \\ &= \left|\mathbf{x}_t(a)^\top\Sigma_W^{-1/2}(t)\left(\tilde{w}(t) - \hat{W}(t)\theta_t\right)\right| \\ &\le v_t^2\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\left\|\frac{1}{v_t^2}\Sigma_W(t)^{1/2}\left(\tilde{w}(t) - \hat{W}(t)\theta_t\right)\right\|_2 \\ &\le v_t^2\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\sqrt{4m\ln t}\end{aligned}$$

with probability $1-\frac{1}{t^2}$. Taking a union bound over all $i \in [d]$, we have with probability $1-\dfrac{1}{t^2}$ for all $j \in [d], a \in [K]$

$$\left|\tilde{w}(t)^\top\mathbf{x}_t(a) - \theta_t^\top\hat{W}(t)^\top\mathbf{x}_t(a)\right| \le v_t^2\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}\sqrt{4m\log dt}$$

Further, using Lemma 6 from (Agrawal & Goyal, 2013), taking a union bound over $j \in [d]$, we have with probability $1-\frac{1}{t^2}$ for all $j \in [d], a \in [K]$:

$$\left|\theta_t^\top\tilde{W}(t)(t)_j^\top\mathbf{x}_t(a) - \theta_t^\top\hat{W}_t(t)_j\mathbf{x}_t(a)\right| \le \sqrt{4\log(tKd)}\; v_t^2\;\sqrt{\mathbf{x}_t(a)^\top\Sigma(t)^{-1}\mathbf{x}_t(a)}$$

Combining with the previous bound completes the proof. □

Now we define the set of saturated actions at time t as

$$\mathcal{C}(t) = \left\{ a \in [K] : \ell_t(a_t^*) - \ell_t(a) > g_t \sqrt{\mathbf{x}_t(a)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a)} \right\}.$$

490 Let $\hat{a}_t$ denote the unsaturated arm with smallest $\sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)}$ in $\mathcal{C}(t)$ i.e.,

$$\hat{a}_t = \operatorname*{argmin}_{a \in \mathcal{C}(t)} \sqrt{\mathbf{x}_t(a)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a)}.$$

491 Using this we have,

$$\begin{aligned}
\ell(a_t^*) - \ell(a_t) &= \mathbf{x}_t(a_t^*)^\top \mu_* - Z\theta_t^\top \left(\frac{B}{T} \mathbb{1} - W_*^\top \mathbf{x}_t(a_t^*) \right) - \mathbf{x}_t(a_t)^\top \mu_* + Z\theta_t^\top \left(\frac{B}{T} \mathbb{1} - W_*^\top \mathbf{x}_t(a_t) \right) \\
&= \mathbf{x}_t(a_t^*)^\top \mu_* - \mathbf{x}_t(a_t)^\top \mu_* + Z\theta_t^\top \Big(W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(a_t) \Big) \\
&= \mathbf{x}_t(a_t^*)^\top \mu_* - \mathbf{x}_t(a_t)^\top \mu_* + Z\theta_t^\top \Big(W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(a_t) \Big) \\
&= \mathbf{x}_t(a_t^*)^\top \mu_* - \mathbf{x}_t(\hat{a}_t)^\top \mu_* + \mathbf{x}_t(\hat{a}_t)^\top \mu_* - \mathbf{x}_t(a_t)^\top \mu_* \\
&\quad + Z\theta_t^\top \Big(W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(\hat{a}_t) + W_*^\top \mathbf{x}_t(\hat{a}_t) - W_*^\top \mathbf{x}_t(a_t) \Big)
\end{aligned}$$

492 Therefore with $g_t = v_t^2 \min\left\{ \sqrt{4m \log dt}, \sqrt{4 \log(tK)} \right\} \sqrt{\mathbf{x}_t(i)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(i)}$ we have

$$\begin{aligned}
\ell(a_t^*) - \ell(a_t) &\overset{(a)}{\leq} \mathbf{x}_t(a_t^*)^\top \mu_* - \mathbf{x}_t(\hat{a}_t)^\top \mu_* + \mathbf{x}_t(\hat{a}_t)^\top \tilde{\mu}(t) - \mathbf{x}_t(a_t)^\top \tilde{\mu}(t) \\
&\quad + g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + g_t \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)} \\
&\quad + Z\theta_t^\top \Big(W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(\hat{a}_t) \Big) + Z\tilde{w}(t)^\top \mathbf{x}_t(\hat{a}_t) - Z\tilde{w}(t)^\top \mathbf{x}_t(a_t) \\
&\quad + g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + g_t \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)}
\end{aligned}$$

where (a) follows if we assume that E_1^μ, E_2^μ, E_1^W and E_2^W hold true. By our choice of action a_t in line 6 of Algorithm 1 we have

$$(\mathbf{x}_t(\hat{a}_t)^\top \tilde{\mu}(t) + Z\tilde{w}(t)^\top \mathbf{x}_t(\hat{a}_t)) - (\mathbf{x}_t(a_t)^\top \tilde{\mu}(t) + Z\tilde{w}(t)^\top \mathbf{x}_t(a_t)) \leq 0.$$

493 Further observe that

$$\begin{aligned}
Z\theta_t^\top \Big(W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(\hat{a}_t) \Big) &\leq Z\|\theta_t\|_1 \|W_*^\top \mathbf{x}_t(a_t^*) - W_*^\top \mathbf{x}_t(\hat{a}_t)\|_\infty \\
&\leq Z g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)}.
\end{aligned}$$

494 Therefore we have

$$\begin{aligned}
\ell(a_t^*) - \ell(a_t) &\leq 2g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + g_t \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)} \\
&\quad + (Z+1) g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + g_t \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)},
\end{aligned}$$

495 which implies

$$\begin{aligned}
&\mathbb{E}[\ell(a_t^*) - \ell(a_t) | \mathcal{F}_{t-1}] \leq \mathbb{E}\Big[2g_t \sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + g_t \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)} | \mathcal{F}_{t-1} \Big] \\
&+ P\big(\{E_2^\mu\}^c\big) + (Z+1) g_t \mathbb{E}\Big[\sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(\hat{a}_t)} + \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1} \mathbf{x}_t(a_t)} | \mathcal{F}_{t-1} \Big] \\
&\qquad + P\big(\{E_2^W\}^c\big)
\end{aligned}$$

Using Lemma 4 from Agrawal & Goyal (2013) we have

$$\mathbb{E}\Big[\sqrt{\mathbf{x}_t(\hat{a}_t)^\top \hat{\Sigma}(t)^{-1}\mathbf{x}_t(\hat{a}_t)}|\mathcal{F}_{t-1}\Big] \leq \frac{1}{(p-1/t^2)}\mathbb{E}\Big[\sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1}\mathbf{x}_t(a_t)}|\mathcal{F}_{t-1}\Big].$$

Therefore

$$\mathbb{E}[\ell(a_t^*) - \ell(a_t)|\mathcal{F}_{t-1}] \leq \frac{(Z+4)g_t}{(p-1/t^2)}\mathbb{E}\Big[\sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t)^{-1}\mathbf{x}_t(a_t)}|\mathcal{F}_{t-1}\Big] + \frac{4g_t}{pt^2}$$

□

Lemma 3.2. *Suppose Assumption 3.1 holds. Then, with probability at least $1-\delta$, we have*

$$(i) \quad \sum_{t=1}^{\tau}\sum_{a\in[K]} \pi^*(a)\ell_t(a) \geq \tau\frac{OPT}{T} - (Z+2)\sqrt{T\log\frac{2}{\delta}}\,,$$

$$(ii) \quad \sum_{t=1}^{\tau}\ell_t(a_t) \leq \sum_{t=1}^{\tau}\left[\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t^\top\Big(\frac{B}{T}\mathbb{1} - \mathbf{v}_t(a_t)\Big)\right] + 2Z\sqrt{T\log(1/\delta)}\,.$$

Proof. Consider the lagrangian

$$\ell_t(a_t) = \mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t\Big(\frac{B}{T}\mathbb{1} - W_*^\top\mathbf{x}_t(a_t)\Big).$$

Using Assumption 3.1, we have $\mathbb{E}[\mathbf{v}_t(a_t)|\mathbf{x}_t(a_t), \mathcal{H}_{t-1}] = W_*^\top\mathbf{x}_t(a_t)$ and therefore

$$\mathcal{M}_t = Z\theta_t^\top\Big(\mathbf{v}_t(a_t) - W_*^\top\mathbf{x}_t(a_t)\Big)$$

is a martingale difference sequence with respect to the filtration $\mathcal{F}_t$. Further

$$\begin{aligned}|\mathcal{M}_t| &\leq Z\|\theta_t\|_1\|W_*^\top\mathbf{x}_t(a_t)\|_\infty + Z\|\theta_t\|_1\|\mathbf{v}_t(a_t)\|_\infty \\ &\leq 2Z\end{aligned}$$

Using Azuma Hoeffding we have

$$P\left(\sum_{t=1}^{\tau}\mathcal{M}_t > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{4\tau Z^2}\right),$$

and therefore with probability $1-\delta$

$$\sum_{t=1}^{\tau}\ell_t(a_t) \leq \sum_{t=1}^{\tau}\left[\mathbf{x}_t(a_t)^\top\mu_* + Z\theta_t^\top\Big(\frac{B}{T}\mathbb{1} - \mathbf{v}_t(a_t)\Big)\right] + 2Z\sqrt{T\log(1/\delta)}, \tag{12}$$

which completes the proof of part (ii).

Next, note that

$$\sum_{a\in[K]}\pi^*(a)\,\ell_t(a) = \sum_{a\in[K]}\pi^*(a)\,\mathbf{x}_t(a)^\top\mu_* + Z\theta_t^\top\left(\frac{B}{T}\mathbb{1} - \sum_{a\in[K]}\pi^*(a)\,W_*^\top\mathbf{x}_t(a)\right).$$

From our definition of the optimal policy, we have that

$$\frac{\text{OPT}}{T} = \mathbb{E}_{X\sim\mathcal{D}}\,\mu_*^\top X\pi^* + Z\theta_*^\top\left(\frac{B}{T}\mathbb{1} - \mathbb{E}_{X\sim\mathcal{D}}W_*^\top X\pi^*\right)$$

where θ_* is the optimal Lagrange multiplier. We define

$$G_t = \sum_{a\in[K]} \pi^*(a)\mathbf{x}_t(a)^\top \mu_* + Z\theta_t^\top \left(\frac{B}{T}\mathbb{1} - \sum_{a\in[K]} \pi^*(a) W_*^\top \mathbf{x}_t(a) \right)$$
$$- \mathbb{E}_{X\sim\mathcal{D}}\mu_*^\top X\pi^* + Z\theta_*^\top \left(\frac{B}{T}\mathbb{1} - \mathbb{E}_{X\sim\mathcal{D}} W_*^\top X\pi^* \right)$$

Then we have $\mathbb{E}(G_t|\mathcal{F}_{t-1}) \geq 0 \quad \text{and} \quad |G_t| \leq 1 + Z + \dfrac{\text{OPT}}{T} \leq Z + 2$. Using Azuma-Hoeffding, with probability $1-\delta$, we have

$$\sum_{t=1}^{\top} G_t \geq -(z+2)\sqrt{T\log\frac{2}{\delta}},$$

which implies, with probability $1-\delta$,

$$\sum_{t=1}^{\tau}\sum_{a\in[K]} \pi^*(a) l_t(a) \geq \tau\frac{\text{OPT}}{T} - (Z+2)\sqrt{T\log\frac{2}{\delta}},$$

thus completing the proof of part (i).

□

B Regret Bound for LinCCB-TS

Lemma 4.1. *Let Assumptions 4.1 and 4.2 hold. Then, the regret in (5) can be bounded as*

$$\texttt{Reg}(T) \leq \sum_{t\in\mathcal{S}_T} \left(h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \right) + n_T^c \Delta_h, \tag{9}$$

where the set $\mathcal{S}_T$ consists of the rounds until the horizon T when LinCCB-TS *played the TS action and $n_T^c = |\mathcal{S}_T^c|$ is the number of times until T when a baseline action was played.*

Proof. The decomposition follows the same approach as Proposition 2 in (Kazerouni et al., 2017), and we present the proof here for completeness. Recall that $\mathcal{S}_T = \{t \in [T] : a_t = b_t\}$ represents the time steps when the baseline action was selected, while $\mathcal{S}_T^c = \{t \in [T] : a_t = \tilde{a}_t\}$ denotes the time steps when the TS action was chosen. Using the fact that $\mathcal{S}_T \cup \mathcal{S}_T^c = [T]$ we have:

$$\begin{aligned}\texttt{Reg}(T) &= \sum_{t=1}^{T} h(\mathbf{x}_{t,a_t}) - \sum_{t=1}^{T} h(\mathbf{x}_{t,a_t^*}) \\ &= \sum_{t\in\mathcal{S}_T} \left(h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \right) + \sum_{t\in\mathcal{S}_T^c} \left(h(\mathbf{x}_{t,b_t}) - h(\mathbf{x}_{t,a_t^*}) \right)\end{aligned}$$

Using the definition of $\Delta_{b_t}^t = h(\mathbf{x}_{t,b_t}) - h(\mathbf{x}_{t,a_t^*})$ and Assumption 4.2 we get

$$\begin{aligned}\texttt{Reg}(T) &= \sum_{t\in\mathcal{S}_T} \left(h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \right) + \sum_{t\in\mathcal{S}_T^c} \Delta_{b_t}^t \\ &\leq \sum_{t\in\mathcal{S}_T} \left(h(\mathbf{x}_{t,a_t}) - h(\mathbf{x}_{t,a_t^*}) \right) + n_T^c \Delta_h,\end{aligned}$$

□

Lemma 4.2. *Suppose Assumption 4.1 and 4.2 holds. Then, with probability $1-\delta/4$ the number of times the baseline action is played by* `C-SquareCB` *is bounded as*

$$n_T^c \leq \mathcal{O}\left(\frac{m^2 \ \min\left\{m, \log K\right\}\left(\log^2 T + \log T \log(1/\delta)\right)}{\alpha r_l(\Delta_l + \alpha y_l)}\right). \tag{10}$$

Proof. We define $\tau = \max\{1 \leq t \leq T \mid a_t = b_t\}$ to be the last time step when the baseline action was played. From line 6 of Algorithm-2 we have

$$\begin{aligned}
\alpha \sum_{t=1}^{\tau} r_t^*(b_t) &= \sum_{t \in S_{\tau-1}} r_t^*(b_t) - \langle \tilde{\mu}(\tau), \mathbf{x}_\tau(a_\tau)\rangle - \sum_{t \in S_{\tau-1}} \langle \tilde{\mu}(t), \mathbf{x}_t(a_t)\rangle \\
&\quad + \underbrace{\sum_{t \in S_{\tau-1}} \left(\min\left\{\sqrt{2m \ln \frac{t}{\delta}}, \sqrt{2 \ln \frac{tK}{\delta}}\right\} + \sqrt{m \ln \frac{t^3}{\delta}} + 1\right) v_t \sqrt{\mathbf{x}_t(a_t)^T \hat{\Sigma}(t) \mathbf{x}_t(a_t)}}_{(A)} \\
&= \underbrace{\sum_{t \in S_{\tau-1}} \left(r_t^*(b_t) - {\mu_*}^\top \mathbf{x}_t(a_t)\right) + r_{b_\tau}^t - {\mu_*}^\top \mathbf{x}_\tau(a_\tau)}_{I} \\
&\qquad + \underbrace{\sum_{t \in S_{\tau-1}} \left(\mu_* - \tilde{\mu}(t)\right)^\top \mathbf{x}_t(a_t) + \left(\mu_* - \tilde{\mu}(t)\right)^\top \mathbf{x}_\tau(a_\tau)}_{II} \ + \ (A)
\end{aligned}$$

Consider term I:

$$\begin{aligned}
&\sum_{t \in S_{\tau-1}} \left(r_t^*(b_t) - {\mu_*}^\top \mathbf{x}_t(a_t)\right) + r_{b_\tau}^t - {\mu_*}^\top \mathbf{x}_\tau(a_\tau) \\
&\quad = \sum_{t \in S_{\tau-1}} \left(r_t^*(b_t) - {\mu_*}^\top \mathbf{x}_t(a_t^*) + {\mu_*}^\top \mathbf{x}_t(a_t^*) - {\mu_*}^\top \mathbf{x}_t(a_t)\right) \\
&\qquad\qquad + r_{b_\tau}^t - {\mu_*}^\top \mathbf{x}_\tau(a_\tau^*) + {\mu_*}^\top \mathbf{x}_\tau(a_\tau^*) - {\mu_*}^\top \mathbf{x}_\tau(a_\tau) \\
&\quad \leq -(n_\tau + 1)\Delta_\ell + \sum_{t \in S_{\tau-1}} {\mu_*}^\top \mathbf{x}_t(a_t^*) - {\mu_*}^\top \mathbf{x}_t(a_t) + {\mu_*}^\top \mathbf{x}_\tau(a_\tau^*) - {\mu_*}^\top \mathbf{x}_\tau(a_\tau)
\end{aligned}$$

Using proof of Theorem 1 from Agrawal & Goyal (2013) we have with probability $1-\delta$

$$\begin{aligned}
&\sum_{t \in S_{\tau-1}} {\mu_*}^\top \mathbf{x}_t(a_t^*) - {\mu_*}^\top \mathbf{x}_t(a_t) + {\mu_*}^\top \mathbf{x}_\tau(a_\tau^*) - {\mu_*}^\top \mathbf{x}_\tau(a_\tau) \\
&\qquad \leq Cm\sqrt{(n_{\tau-1}+1)} \min\left\{\sqrt{m}, \sqrt{\log K}\right\}\left(\log(n_\tau + 1) + \sqrt{\log(n_\tau+1)\log(1/\delta)}\right)
\end{aligned}$$

for some constant $C > 0$. Therefore term I can be bounded with probability $1-\delta$ as follows

$$\begin{aligned}
&\sum_{t \in S_{\tau-1}} \left(r_t^*(b_t) - {\mu_*}^\top \mathbf{x}_t(a_t)\right) + r_{b_\tau}^t - {\mu_*}^\top \mathbf{x}_\tau(a_\tau) \\
&\qquad \leq Cm\sqrt{(n_{\tau-1}+1)} \min\left\{\sqrt{m}, \sqrt{\log K}\right\} \log(n_\tau + 1) + \sqrt{\log(n_\tau+1)\log(1/\delta)}
\end{aligned}$$

for some constant $C > 0$.

Next consider term II and observe that using Lemma 1 from Agrawal & Goyal (2013) we have with probability $1-\delta$

$$\sum_{t\in S_{\tau-1}} \left(\mu_* - \tilde{\mu}(t)\right)^\top \mathbf{x}_t(a_t) + \left(\mu_* - \tilde{\mu}(t)\right)^\top \mathbf{x}_\tau(a_\tau)$$
$$\leq \sum_{t\in S_{\tau-1}} \left(\sqrt{m\ln\frac{1}{\delta}} + 1 + \min\left\{\sqrt{4m\ln(\frac{1}{\delta})}, \sqrt{4\ln(\frac{K}{\delta})}\right\} v_t\right)\sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)}$$

Combining the bounds on all the terms we get with probability $1-\delta$

$$\alpha\sum_{t=1}^{\tau} r_t^*(b_t) \leq -(n_\tau+1)\Delta_\ell$$
$$+ Cm\sqrt{(n_{\tau-1}+1)}\min\left\{\sqrt{m}, \sqrt{\log K}\right\}\log(n_\tau+1) + \sqrt{\log(n_\tau+1)\log(1/\delta)} + 2\cdot(A)$$

for some constant $C>0$. To bound term (A) defined as

$$(A) = \sum_{t\in S_{\tau-1}} \left(\min\left\{\sqrt{2m\ln\frac{t}{\delta}}, \sqrt{2\ln\frac{tK}{\delta}}\right\} + \sqrt{m\ln\frac{t^3}{\delta}} + 1\right) v_t\sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)}\ ,$$

first note that

$$\sum_{t\in S_{\tau-1}} \left(\min\left\{\sqrt{2m\ln\frac{t}{\delta}}, \sqrt{2\ln\frac{tK}{\delta}}\right\} + \sqrt{m\ln\frac{t^3}{\delta}} + 1\right) v_t\sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)}$$
$$\leq \left(\min\left\{\sqrt{2m\ln\frac{T}{\delta}}, \sqrt{2\ln\frac{TK}{\delta}}\right\} + \sqrt{m\ln\frac{T^3}{\delta}} + 1\right) v_T \sum_{t\in S_{\tau-1}} \sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)}.$$

Using Lemma 3 from Chu et al. (2011) for $t\in\mathcal{S}_{\tau-1}$ we have

$$\sum_{t\in S_{\tau-1}} \sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)} \leq 5\sqrt{m\ n_\tau\ln(n_\tau)}.$$

Therefore with probability $1-\delta$ for some constant $C>0$.

$$\alpha\sum_{t=1}^{\tau} r_t^*(b_t) \leq -(n_\tau+1)\Delta_\ell$$
$$+ Cm\sqrt{(n_{\tau-1}+1)}\min\left\{\sqrt{m}, \sqrt{\log K}\right\}\left(\log(T+1) + \sqrt{\log(T+1)\log(1/\delta)}\right)$$

Now, using the fact that $n_{\tau-1}+n^c_{\tau-1}+1=\tau$, and Assumption 4.2, we have $r_l \leq r_i^*(b_i) \leq r_h, \forall i \in [T]$. Therefore,

$$\alpha\sum_{i=1}^{\tau} r_t^*(b_t) \geq \alpha\ (n_{\tau-1}+n^c_{\tau-1}+1)\ r_l.$$

Therefore, with probability $1-\delta$, we obtain

$$\alpha n^c_{\tau-1} r_l \leq -(n_{\tau-1}+1)(\Delta_l+\alpha r_l)$$
$$+ Cm\sqrt{(n_{\tau-1}+1)}\min\left\{\sqrt{m}, \sqrt{\log K}\right\}\left(\log(T+1) + \sqrt{\log(T+1)\log(1/\delta)}\right)$$

Using $n_T^c = n_{\tau-1} + 1$, with probability $1-\delta$, we have

$$n_T^c \le \frac{1}{\alpha r_l}\Bigg\{ -(n_{\tau-1}+1)(\Delta_l + \alpha r_l) + Cm\sqrt{(n_{\tau-1}+1)}\min\left\{\sqrt{m}, \sqrt{\log K}\right\}\Big(\log(T+1) + \sqrt{\log(T+1)\log(1/\delta)}\Big)\Bigg\}.$$

Following Deb et al. (2025) we define the following function

$$Q(n_{\tau-1}) = \Bigg\{ -(n_{\tau-1}+1)(\Delta_l + \alpha r_l) + Cm\sqrt{(n_{\tau-1}+1)}\min\left\{\sqrt{m}, \sqrt{\log K}\right\}\Big(\log(T+1) + \sqrt{\log(T+1)\log(1/\delta)}\Big)\Bigg\}.$$

Note that $Q(n_{\tau-1}) \le -c_1 n + c_2\sqrt{n} := f(n)$, where

$$\begin{aligned} c_1 &= \Delta_l + \alpha r_l \ge 0, \\ c_2 &= Cm\ \min\left\{\sqrt{m}, \sqrt{\log K}\right\}\Big(\log(T+1) + \sqrt{\log(T+1)\log(1/\delta)}\Big), \\ n &= n_{\tau-1} + 1. \end{aligned}$$

Setting $f'(n) = 0$, and solving we get $n^* = \dfrac{c_2^2}{4c_1^2}$. Therefore,

$$\begin{aligned} Q(m_{\tau-1}) &\le -\frac{c_2^2}{4c_1} + \frac{c_2^2}{2c_1} \\ &= \frac{c_2^2}{4c_1} \\ &\le \mathcal{O}\left(\frac{m^2\ \min\left\{m, \log K\right\}\Big(\log^2 T + \log T\log(1/\delta)\Big)}{\Delta_l + \alpha y_l}\right). \end{aligned}$$

Combining with the upper bound for n_T^c we get with probability $1-\delta$

$$n_T^c \le \mathcal{O}\left(\frac{m^2\ \min\left\{m, \log K\right\}\Big(\log^2 T + \log T\log(1/\delta)\Big)}{\alpha r_l(\Delta_l + \alpha y_l)}\right).$$

□

Lemma 4.4. *Let Assumptions 4.1 and 4.2 hold. Then, for any $\delta > 0$, with probability $1-\delta$,* `LinCCB-TS` *satisfies the performance constraint in* (6).

Proof. Note that for any $t \in [T]$ the *safety condition* ensures that

$$\begin{aligned} &\sum_{\tau\in\mathcal{S}_{t-1}} \tilde{\mu}(\tau)^\top \mathbf{x}_\tau(a_\tau) + \tilde{\mu}(t)^\top \mathbf{x}_t(\tilde{a}_t) + \sum_{\tau\in S_{t-1}^c} r_\tau^*(b_\tau) \\ &\quad - \sum_{t\in S_{\tau-1}}\left(\min\left\{\sqrt{2m\ln\frac{t}{\delta}}, \sqrt{2\ln\frac{tK}{\delta}}\right\} + \sqrt{m\ln\frac{t^3}{\delta}} + 1\right) v_t\sqrt{\mathbf{x}_t(a_t)^T\hat{\Sigma}(t)\mathbf{x}_t(a_t)} \\ &\qquad \ge (1-\alpha)\sum_{\tau=1}^{t} r_\tau^*(b_\tau) \end{aligned}$$

Now using Lemma 1 from Agrawal & Goyal (2013) we have that with probability $1-\delta$ for any $a \in [K]$

$$
\begin{aligned}
&|\tilde{\mu}(\tau)^\top \mathbf{x}_\tau(a_\tau) - \mu^*(\tau)^\top \mathbf{x}_\tau(a_\tau)| \\
&\qquad \leq \left(\sqrt{m \ln \frac{t}{\delta}} + 1 + \min\{ \sqrt{4m\ln(\frac{t}{\delta})}, \sqrt{4\ln(\frac{Kt}{\delta})} \} v_t \right) \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t) \mathbf{x}_t(a_t)}
\end{aligned}
$$

and therefore we have with probability $1-\delta$

$$
\begin{aligned}
&\sum_{t \in S_{\tau-1}} |\tilde{\mu}(\tau)^\top \mathbf{x}_\tau(a_\tau) - \mu^*(\tau)^\top \mathbf{x}_\tau(a_\tau)| \\
&\leq \sum_{t \in S_{\tau-1}} \left(\sqrt{m \ln \frac{t^3}{\delta}} + 1 + \min\left\{ \sqrt{4m\ln\left(\frac{t}{\delta}\right)}, \sqrt{4\ln\left(\frac{Kt}{\delta}\right)} \right\} v_t \right) \sqrt{\mathbf{x}_t(a_t)^\top \hat{\Sigma}(t) \mathbf{x}_t(a_t)}
\end{aligned}
$$

Combining with the safety condition we have with probability $1-\delta$

$$
\begin{aligned}
\sum_{\tau \in \mathcal{S}_{t-1}} \mu^*(\tau)^\top \mathbf{x}_\tau(a_\tau) + \mu^*(t)^\top \mathbf{x}_t(\tilde{a}_t) + \sum_{\tau \in S^c_{t-1}} r^*_\tau(b_\tau) \\
\geq (1-\alpha) \sum_{\tau=1}^{t} r^*_\tau(b_\tau)
\end{aligned}
$$

which implies

$$
\sum_{\tau=1}^{t} r^*_\tau(a_\tau) \geq (1-\alpha) \sum_{\tau=1}^{t} r^*_\tau(b_\tau)
$$

thus completing the proof. □