Smoothed Adversarial Linear Contextual Bandits with Knapsacks

Vidyashankar Sivakumar [1] **Shiliang Zuo** [2] **Arindam Banerjee** [2]

Abstract

Many bandit problems are characterized by the learner making decisions under constraints. The learner in Linear Contextual Bandits with Knapsacks (LinCBwK) receives a resource consumption vector in addition to a scalar reward in each time step which are both linear functions of the context corresponding to the chosen arm. For a fixed time horizon T, the goal of the learner is to maximize rewards while ensuring resource consumptions do not exceed a pre-specified budget. We present algorithms and characterize regret for LinCBwK in the smoothed setting where base context vectors are assumed to be perturbed by Gaussian noise. We consider both the stochastic and adversarial settings for the base contexts, and our analysis of stochastic LinCBwK can be viewed as a warm-up to the more challenging adversarial LinCBwK. For the stochastic setting, we obtain $O(\sqrt{T})$ additive regret bounds compared to the best context dependent fixed policy. The analysis combines ideas for greedy parameter estimation in (Kannan et al., 2018; Sivakumar et al., 2020) and the primal-dual paradigm first explored in (Agrawal & Devanur, 2016; 2014a). Our main contribution is an algorithm with $O(\log T)$ competitive ratio relative to the best context dependent fixed policy for the adversarial setting. The algorithm for the adversarial setting employs ideas from the primal-dual framework (Agrawal & Devanur, 2016; 2014a) and a novel adaptation of the doubling trick (Immorlica et al., 2019).

1. Introduction

The contextual bandits framework (Langford & Zhang, 2007; Lattimore & Szepesvari, 2019; Slivkins, 2021) is a popular sequential decision making framework used in multiple practical applications such as clinical trials, web search, and content optimization. Contextual bandit problems have multiple rounds where in each round a learner watches a set of context vectors corresponding to K arms and chooses an arm with the goal of maximizing cumulative rewards. In linear contextual bandits (LinCB) (Chu et al., 2011; Li et al., 2010), the rewards for an arm are a linear function of the context vector. Algorithms for contextual bandit problems typically have to balance between *exploration*, choosing potentially sub-optimal arms for acquiring more information, and *exploitation*, choosing arms to optimize immediate rewards. Motivated by fairness and ethics considerations (Bird et al., 2016; Raghavan et al., 2018; Kannan et al., 2018) or to avoid inefficient exploration strategies (Bastani et al., 2018), recent work has studied settings where an exploration free greedy algorithm can be employed. The crux of all such prior work is an assumption of inherent randomness in context vectors aiding exploration. In the smoothed bandits framework (Kannan et al., 2018) the inherent randomness is due to stochastic perturbations of the context vectors.

Often a learner has to operate under constraints while maximizing rewards (Babaioff et al., 2015; Badanidiyuru et al., 2012; Besbes & Zeevi, 2012; Singla & Krause, 2013; Combes et al., 2015). For example clinical trials are constrained by available medical resources or optimizing for ad placements should account for advertisers budget and user reach. There is now a body of work under the theme *bandits with knapsacks (BwK)* addressing the tension between maximizing rewards while satisfying constraints (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014a; 2016; Immorlica et al., 2019; Agrawal et al., 2016; Badanidiyuru et al., 2014).

Smoothed LinCBwK. In this work, we consider the linear contextual bandits with knapsacks (LinCBwK) problem (Agrawal & Devanur, 2016) under the smoothed assumption on context vectors. The LinCBwK problem has d resources with budget $B \in \mathbb{R}_+$ for each resource. In each round t, there are K context vectors $\{\mathbf{x}_t(a)\}_{a=1}^K, \mathbf{x}_t(a) \in \mathbb{R}^m$, one corresponding to each arm $a \in [K]$. The smoothed context vectors are of the form $\mathbf{x}_t(a) = \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a)$,

[1] Amazon [2] Department of Computer Science, University of Illinois, Urbana-Champaign. Correspondence to: Vidyashankar Sivakumar <shankar861@gmail.com>, Shiliang Zuo <szuo3@illinois.edu>, Arindam Banerjee <arindamb@illinois.edu>. All work by VS done prior to starting employment with Amazon.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

where $\boldsymbol{\nu}_t(a) \in \mathbb{B}_2^m$ (unit ball) is the base context vector and $\mathbf{g}_t(a) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_{m \times m})$ are Gaussian perturbations. Typically $\sigma^2 = \frac{1}{m}$ (Kannan et al., 2018; Sivakumar et al., 2020). On choosing an arm $a_t \in [K]$, the learner receives a noisy reward $r_t(a_t) \in \mathbb{R}_+$ and consumption vector $\mathbf{v}_t(a_t) \in \mathbb{R}_+^d$ satisfying

$$\mathbb{E}[r_t(a) \mid \mathbf{x}_t(a), H_{t-1}] = \boldsymbol{\mu}_*^\intercal \mathbf{x}_t(a) \ , \qquad (1)$$

$$\mathbb{E}[\mathbf{v}_t(a) \mid \mathbf{x}_t(a), H_{t-1}] = W_*^\intercal \mathbf{x}_t(a) \ , \qquad (2)$$

where H_{t-1} denotes the history, the reward vector $\boldsymbol{\mu}_* \in S^{m-1}$ (unit sphere) and the consumption matrix $W_* \in \mathbb{R}^{m \times d}$ with columns in S^{m-1} are fixed but unknown to the learner. Additionally, the parameter $\boldsymbol{\mu}_*$ and columns of W_* can be assumed to have some structure like sparsity. In each round, the learner also has a "no-op" option which is to choose none of the arms and receive 0 reward with $\mathbf{0}$ resource consumption. The learner's goal is to choose arms which maximize the total reward over T time steps while ensuring total consumption for each resource does not exceed B.

$$\max \ \sum_{t=1}^{T} r_t(a_t) \qquad \text{s.t.} \qquad \sum_{t=1}^{T} \mathbf{v}_t(a_t) \leqslant \mathbf{1}B \ . \qquad (3)$$

Stochastic and Adversarial Contexts. We study algorithms under two assumptions on the base context vectors: (a) Stochastic LinCBwK when the base context vectors $\boldsymbol{\nu}_t(a)$ are sampled from an unknown distribution, and (b) Adversarial LinCBwK when the $\boldsymbol{\nu}_t(a)$ are chosen by an adaptive adversary. In both cases, we compare the learner's performance against an *optimal adaptive policy* with knowledge of $\boldsymbol{\mu}_*, W_*$ which, unlike LinCB, is no longer a single arm but a probability distribution over the arms (Agrawal & Devanur, 2014a; Badanidiyuru et al., 2013; Immorlica et al., 2019). Compared to LinCB, the learner's algorithm is more complicated because it should not only optimize the per step reward but also account for how much resources are being consumed vs. how much to conserve for future.

The tension between consumption vs. conservation differentiates the algorithms for stochastic and adversarial LinCBwK. While in the stochastic setting, historical data can be extrapolated to guide decisions to consume vs. conserve, it is impossible in the adversarial setting to use historical data to plan for the future. Thus, for stochastic LinCBwK, we are able to achieve stronger additive regret bounds with respect to the optimal adaptive policy. Let OPT_1 denote the optimal adaptive policy's reward for smoothed stochastic LinCBwK. We present an algorithm whose reward REW satisfies

$$\mathrm{REW} \geqslant \mathrm{OPT}_1 - \tilde{O}\left(\left(\frac{\mathrm{OPT}_1}{B} + 1\right)\frac{\sqrt{T}}{\sigma}\right) \ , \qquad (4)$$

where $\tilde{O}(\cdot)$ notation hides dependence on logarithmic factors and dimension of the problem.

For adversarial LinCBwK, we are only able to bound the competitive ratio, i.e., the ratio between the optimal adaptive policy's reward and the algorithm's reward. With OPT_2 denote the optimal adaptive policy's reward for the smoothed adversarial LinCBwK, we present an algorithm whose reward REW satisfies

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}_2}{O(d\lceil \log T \rceil)} - \tilde{O}\left(\left(\frac{\mathrm{OPT}_2}{B} + 1\right)\frac{\sqrt{T}}{\sigma}\right) \ . \quad (5)$$

Our framework is general enough to handle structure assumptions like sparsity on the parameter vectors.

LinCBwK Algorithms. Algorithms for both stochastic and adversarial LinCBwK broadly perform two steps in each round: 1) estimating reward and consumptions for each arm, and 2) pulling arms while balancing earned reward and resource consumptions. In the smoothed setting, we use (episodic) greedy estimates of reward and consumptions for each arm obtained using constrained least squares estimates of the reward and consumption parameters (Kannan et al., 2018; Sivakumar et al., 2020). This is in sharp contrast with most existing work in contextual bandits including prior work on stochastic LinCBwK (Agrawal & Devanur, 2016) which compute rewards based on upper confidence bound (UCB). For choosing arms by balancing the tradeoff between reward and consumptions, multiple primal-dual approaches have been explored in BwK (Badanidiyuru et al., 2013; Agrawal & Devanur, 2016; 2014b; Immorlica et al., 2019) and related literature on online stochastic packing problems (Agarwal et al., 2014; Mehta, 2013; Mehta et al., 2007; Buchbinder & Naor, 2009; Williamson & Shmoys, 2011; Devanur & Hayes, 2009; Agrawal & Devanur, 2014b; Devanur et al., 2011; Feldman et al., 2010; Molinaro & Ravi, 2012). Our algorithm is built on the framework developed in (Agrawal & Devanur, 2014b; 2016) where in each round, the algorithm maximizes a linear combination of the reward and consumptions with a tradeoff parameter Z. The function can be viewed as the Lagrangian of the constrained linear program (LP) in (3) with suitable modifications due to the sequential and bandit nature of the problem. Z can be viewed as a global parameter which captures the tradeoff between the optimal reward and the budget. Specifics of the computation of Z differs for stochastic and adversarial LinCBwK, being substantially more challenging for the adversarial setting. Our algorithms provide ways of computing Z and comes with regret bounds as in (4), (5).

Comparison with Prior Work. We analyze smoothed LinCBwK where the (stochastic or adversarial) contexts are perturbed by a small amount of Gaussian noise. The smoothing assumption (Kannan et al., 2018; Sivakumar et al., 2020) is a middle ground between assuming stochastic independent contexts and adversarial contexts both of which arguably are not representative of most real world problems. In sharp contrast to the explore-exploit bandit algorithms

Table 1. Comparisons with prior work on regret upper bounds and budget constraints. OPT is an upper bound on optimal reward, K is number of arms, B is the budget, m is context dimension for the LinCBwK problem, d is number of constraints.

Setting	Regret	Budget B
Stochastic BwK (Badanidiyuru et al., 2013)	$O(\sqrt{K}\,\mathrm{OPT}(1+\sqrt{\mathrm{OPT}/B}))$	
Adversarial BwK (Immorlica et al., 2019)	$\mathrm{OPT}/O(d\log T) - O(T^{7/4}K/B)$	$\Omega(KT^{3/4})$
Stochastic LinCBwK (Agrawal & Devanur, 2016)	$O(m\sqrt{T})$	$\Omega(mT^{3/4})$
Smoothed Stochastic LinCBwK (This paper)	$O(m\sqrt{T})$	$\Omega(m^{2/3}T^{3/4})$
Smoothed Adversarial LinCBwK (This paper)	$\mathrm{OPT}/O(d\log T) - O(m\sqrt{T})$	$\Omega(T^{3/4})$

in prior literature (Agrawal & Devanur, 2016; Immorlica et al., 2019), we show an *exploration free greedy algorithm* theoretically achieves optimal regret in the smoothed setting. Practically such exploration free bandit algorithms are desirable in problems where ethics, fairness (Bird et al., 2016), or computational efficiency (Bastani et al., 2018) are concerns. For LinCB, (Bietti et al., 2018) empirically show the greedy algorithm to perform as well as state-of-the-art explore-exploit bandit algorithms on many practical datasets.

Table 1 is a summary of prior results for variations of the bandit with knapsacks problems. Our results for stochastic LinCBwK match upto log factors the regret bounds for LinCBwK in (Agrawal & Devanur, 2016) and the bounds for LinCB with smoothed contexts in (Sivakumar et al., 2020). Our main contribution is an algorithm and regret analysis for smoothed adversarial LinCBwK. Although we do not rigorously analyze lower bounds, our algorithm achieves worst case competitive ratio which match upto constant factors the bounds in (Immorlica et al., 2019) for the arguably simpler setting of adversarial multi armed bandits (MAB) with knapsacks. We also analyze regret bounds in the high-dimensional regularized setting, e.g., Lasso.

Notations: Vectors are denoted by bold symbols, e.g., $\boldsymbol{\mu}, \mathbf{y}$, and matrices are denoted by upper case letters, e.g., W, X. For context $\mathbf{x}_t(a) = \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a), a \in [K]$, we will use $X_t, N_t, G_t \in \mathbb{R}^{m\times(K+1)}$ with $X_t = N_t + G_t$ to denote the set of base context vectors, Gaussian perturbation vectors, and observed context vectors respectively in round t. The last column in X_t, N_t, G_t is $\mathbf{0}$ corresponding to the no-op arm. $\mathbf{r}_t \in \mathbb{R}^{K+1}$ and $V_t \in \mathbb{R}^{d\times(K+1)}$ will denote the reward and consumption vectors at time t.

2. Smoothed LinCBwK

As discussed in Section 1, in the smoothed setting, the context vectors are Gaussian perturbed versions of base context vectors, i.e., for all $a \in [K]$,

$$\mathbf{x}_t(a) = \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a)\ , \tag{6}$$

where $\boldsymbol{\nu}_t(a) \in \mathbb{B}_2^m$ are the base context vectors and $\mathbf{g}_t(a) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}_{m\times m})$ are Gaussian perturbations. We consider two settings for the base context vectors $\boldsymbol{\nu}_t(a)$: (a) stochastic, where the $\boldsymbol{\nu}_t(a)$ are stochastically chosen independently from a fixed but unknown distribution, and (b) adversarial, where the $\boldsymbol{\nu}_t(a)$ are chosen by an adaptive adversary. The following result establishes high probability bounds on $\|\mathbf{x}_t(a)\|_2$.

Lemma 1 *Let* $\beta := \operatorname*{argmax}_{t\in[T],a\in[K]} \|\mathbf{x}_t(a)\|_2$. *Then with probability at least* $(1-\delta)$,

$$\beta \leqslant O\left(1 + \sigma\sqrt{m\log(TK/\delta)}\right)\ . \tag{7}$$

We will assume $\sigma^2 = 1/m$. Since the parameter vectors have L_2 norm unity the rewards and consumptions in each time step are bounded by $O(\beta)$.

Our framework can handle sparsity assumptions on $\boldsymbol{\mu}_*$ and columns of W_*. Without loss of generality, we will assume the same structure for $\boldsymbol{\mu}_*$ and the columns of W_*. The sparsity structure is represented by an atomic norm $J(\cdot)$ (Chandrasekaran et al., 2012). We use the constrained least squares estimator in each step for estimation.

$$\hat{\boldsymbol{\mu}}_t = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathbb{R}^m} \frac{1}{2t}\|\tilde{\mathbf{y}}_t - \tilde{S}_t\boldsymbol{\mu}\|_2^2 \quad \text{s.t.} \quad J(\boldsymbol{\mu}) \leqslant J(\boldsymbol{\mu}_*)\ ,$$

and

$$(\widehat{W}_t)_{:j} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^m} \frac{1}{2t}\|(\tilde{Q}_t)_{:j} - \tilde{S}_t\mathbf{w}\|_2^2 \quad \text{s.t.} \quad J(\mathbf{w}) \leqslant J(W_{*j}),$$

where $\tilde{S}_t \in \mathbb{R}^{t\times m}$ is the matrix with contexts chosen before time t as rows, $\tilde{\mathbf{y}}_t \in \mathbb{R}^t$ are the rewards observed before time t and $\tilde{Q}_t \in \mathbb{R}^{t\times d}$ are the consumption vectors observed before time t. Let $E_c := \{\mathbf{e} \mid J(\boldsymbol{\mu}_* + \mathbf{e}) \leqslant J(\boldsymbol{\mu}_*)\}$ denote the error set and $A = \text{cone}\,(E_c) \cap S^{m-1}$ denote the error cone (Wainwright, 2019; Vershynin, 2018; Banerjee et al., 2014). Error sets for the columns of W_* are defined similarly. Our results for regret will have dependence on the Gaussian width of the error set $w(A)$ (Talagrand, 2005; 2014). For example $w(A) = \Theta(\sqrt{m})$ if no structure is assumed for the parameter vectors and $w(A) = \Theta(\sqrt{s\log m})$ if the parameters are s-sparse and Lasso (Tibshirani, 1996) is used for parameter estimation. Henceforth we will overload the term A to denote the error cone for constrained estimation of $\boldsymbol{\mu}_*$ and columns of W_*.

We will compare our algorithms to the class of dynamic policies. Formally, let $\phi(T) = (X_t, \mathbf{r}_t, V_t)_{t=1}^T$ denote a particular instantiation of contexts for T time steps. $\mathbf{r}_t \in \mathbb{R}^{K+1}, V_t \in \mathbb{R}^{d\times(K+1)}$ denote the reward and consumption vectors at time t including for the no-op arm. A dynamic policy $p_t : \phi(T) \mapsto \Delta_{K+1}$ maps contexts at each time step to distributions over $K+1$ options, viz. the K arms and the no-op option. In particular, we develop regret bounds with respect to the *optimal dynamic policy* $p^*(\cdot)$ which has foreknowledge of $\boldsymbol{\mu}_*, W_*$ and the mechanism by which the contexts are generated. Interestingly, we will in fact work with a suitably defined context dependent static policy as a benchmark by showing that its reward upper bounds the reward of the optimal dynamic policy while satisfying the resource constraints.

Benchmarks: Stochastic setting. In the stochastic setting X_t's are generated from an unknown distribution $\mathcal{D}$. Here $\mathcal{D}$ is the distribution for the convolution $X_t = N_t + G_t$ with $X_t, N_t, G_t \in \mathbb{R}^{m\times(K+1)}$ denoting the observed context matrix, base vectors matrix and Gaussian perturbation matrix respectively. The last column of all matrices is the $\mathbf{0}$ vector corresponding to the no-op arm. Let $\pi : \mathbb{R}^{m\times(K+1)} \mapsto \Delta_{K+1}$ denote a context dependent probability distribution over arms and let $r(\pi)$ and $\mathbf{v}(\pi)$ denote the expected reward and consumption vector of policy $\pi(\cdot)$, i.e.,

$$r(\pi) := \mathbb{E}_{X\sim\mathcal{D}}[\mathbf{r}^\intercal \pi(X;\mathcal{D})], \quad \mathbf{v}(\pi) := \mathbb{E}_{X\sim\mathcal{D}}[V\pi(X;\mathcal{D})]\,.$$

Let,

$$\pi^* := \operatorname*{argmax}_{\pi} \; T\mathbf{r}(\pi) \quad \text{s. t.} \quad T\mathbf{v}(\pi) \leqslant B\mathbf{1}\,. \tag{8}$$

We revisit a result by (Agrawal & Devanur, 2016) which shows that in the stochastic setting, the reward obtained by the optimal dynamic policy $p^*(\cdot)$ can be upper bounded by an optimal context dependent static policy $\pi^*(\cdot)$.

Lemma 2 (Lemma 1 in (Agrawal & Devanur, 2016)) *Let $\overline{\mathrm{OPT}}_1 := \mathbb{E}_{\phi(T)}[\sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t p_t^*(\phi(T))]$ denote the value of the optimal feasible dynamic policy p^* which knows parameters μ_*, W_* and distribution $\mathcal{D}$. Then, the optimal static policy π^* in* (8) *satisfies: $Tr(\pi^*) \geqslant \overline{\mathrm{OPT}}_1$ and $Tv(\pi^*) \leqslant B\mathbf{1}$.*

Since the optimal static policy π^* is competitive with the optimal dynamic policy p^*, for the stochastic setting it suffices to compare the performance of our proposed algorithm with the performance of π^*. If the sequence of arms chosen by our algorithm is $\{a_t\}$, we define regret as: $R_1(T) := \mathrm{OPT}_1 - \sum_{t=1}^T \mathbf{r}_t(a_t)$, where $\mathrm{OPT}_1 := Tr(\pi^*)$.

Benchmarks: Adversarial setting. In the smoothed adversarial setting, the base context vectors $\boldsymbol{\nu}_t(a), a \in [K]$ are chosen by an adaptive adversary, and the optimal feasible dynamic policy p^* is the one which maximizes the cumulative reward $p^* := \operatorname{argmax}_p \sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t p_t(\phi(T))$ while staying feasible. As in the smoothed stochastic setting, we consider the optimal static policy π^* that depends on context X_t and realization $\phi(T)$. The policy π^* optimizes the cumulative reward while staying feasible, i.e.,

$$\begin{aligned} \pi^* := \operatorname*{argmax}_{\pi} & \sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t \pi(X_t;\phi(T)) \\ \text{s. t.} & \sum_{t=1}^T W_*^\intercal X_t \pi(X_t;\phi(T)) \leqslant B\mathbf{1}\,. \end{aligned} \tag{9}$$

Lemma 3 *Let $\overline{\mathrm{OPT}}_2 := \sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t p_t^*(\phi(T))$ and $\mathrm{OPT}_2 := \sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t \pi^*(X_t;\phi(T))$ respectively denote the value of the optimal feasible dynamic and static policies which have knowledge of the parameters $\boldsymbol{\mu}_*, W_*$. Then $\mathrm{OPT}_2 \geqslant \overline{\mathrm{OPT}}_2$.*

Since the optimal static policy again dominates the optimal dynamic policy under feasibility, it suffices to compare the performance of our algorithm to that of the static policy π^*. If the sequence of arms chosen by our algorithm is $\{a_t\}$, we define regret as: $R_2(T) := \mathrm{OPT}_2 - \sum_{t=1}^T \mathbf{r}_t(a_t)$, where $\mathrm{OPT}_2 := \sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t \pi^*(X_t;\phi(T))$.

3. Greedy Algorithm

In this section we discuss the greedy algorithm `GreedyLinCBwK` (Algorithm 1) used in both the stochastic and adversarial settings. `GreedyLinCBwK` is based on the primal-dual paradigm explored in prior work on BwK (Agrawal & Devanur, 2014a;b; 2016; Immorlica et al., 2019) with a key unique aspect: our algorithm keeps updating estimates $\widehat{\boldsymbol{\mu}}_t, \widehat{W}_t$ respectively of the true reward and consumption parameters $\boldsymbol{\mu}_*, W_*$ and picks arms greedily using these estimates. The greedy approach of arm selection in `GreedyLinCBwK` is in sharp contrast with both the existing UCB-based approach for LinCBwK in the stochastic setting (Agrawal & Devanur, 2014a;b; 2016) and the approach for adversarial BwK setting in (Immorlica et al., 2019). The overall approach in `GreedyLinCBwK` has similarities with prior work on both stochastic and adversarial bandits with knapsacks (Agrawal & Devanur, 2016; Immorlica et al., 2019) in that the primal arm selection is with bandit feedback and the dual update is essentially a full information online convex optimization (OCO).

Algorithm 1 `GreedyLinCBwK`

```
  Input: Parameter $Z \in \mathbb{R}_+$, budget $B$, context matrix $S_1$, reward vector $\mathbf{y}_1$, consumption matrix $Q_1$
  Initialize $\boldsymbol{\theta}_1$ using uniform distribution
  for $t = 1$ to $T$ do
    $(S, \mathbf{y}, Q, \boldsymbol{\theta}, T_{\text{eff}})_{t+1} = \texttt{GreedyStep}(Z, B, (S, \mathbf{y}, Q, \boldsymbol{\theta}, T_{\text{eff}})_t)$
    if $Q_{t+1}\mathbf{e}_j \geqslant B$ for any $j \in [d]$ then
      EXIT
    end if
  end for
```

Algorithm 2 `GreedyStep`

```
  Input: Parameter $Z \in \mathbb{R}_+$, budget $B$, context matrix $S_t$, reward vector $\mathbf{y}_t$, consumption matrix $Q_t$, dual vector $\theta_t$, $T_{\text{eff}}(t)$
  Estimate $\widehat{\boldsymbol{\mu}}_t, \widehat{W}_t = \texttt{Estimate}(S_t, \mathbf{y}_t, Q_t, T_{\text{eff}}(t))$
  Set $a_t = \operatorname*{argmax}_{a \in [K]} \ \mathbf{x}_t(a)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t)$ .
  if $\mathbf{x}_t(a_t)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) \geqslant -2\zeta_t$ as in (10) then
    Select arm $a_t$, $T_{\text{eff}}(t+1) = T_{\text{eff}}(t) + 1$
    Observe reward $r_t(a_t)$ and consumption $\mathbf{v}_t(a_t) \in \mathbb{R}^d$
    Append $\mathbf{x}_t(a_t)$, $r_t(a_t)$, $\mathbf{v}_t(a_t)$ to $S_t$, $\mathbf{y}_t$, $Q_t$ to get $S_{t+1}, \mathbf{y}_{t+1}, Q_{t+1}$.
  else
    Set $a_t =$ the no-op arm, $T_{\text{eff}}(t+1) = T_{\text{eff}}(t)$
    Set $S_{t+1}, \mathbf{y}_{t+1}, Q_{t+1} = S_t, \mathbf{y}_t, Q_t$.
  end if
  Update $\boldsymbol{\theta}_{t+1}$ using OMD with $g_t(\boldsymbol{\theta}) := \langle \boldsymbol{\theta}, (\mathbf{v}_t(a_t) - \frac{B_0}{T}\mathbf{1}) \rangle$ and $\boldsymbol{\theta}_t \in \Delta_{d+1}$
```

Algorithm 3 `Estimate`

```
  Input: Context matrix $S_t$, reward vector $\mathbf{y}_t$, consumption matrix $Q_t$, effective steps $T_{\text{eff}}(t)$
  Compute SVD of design matrix: $\frac{1}{\sqrt{T_{\text{eff}}(t)}} S_t = UDV^\intercal$
  Compute Puffer transformation: $F = UD^{-1}U^\intercal$ and define $\tilde{S}_t = FS_t$, $\tilde{\mathbf{y}}_t = F\mathbf{y}_t$, $\tilde{Q}_t = FQ_t$
  Estimate parameters using constrained least squares

  $\widehat{\boldsymbol{\mu}}_t = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^m} \frac{1}{2T_{\text{eff}}(t))} \|\tilde{\mathbf{y}}_t - \tilde{S}_t\boldsymbol{\mu}\|_2^2 \ \text{ s.t. } J(\boldsymbol{\mu}) \leqslant J(\boldsymbol{\mu}_*)$

  $(\widehat{W}_t)_{:j} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2T_{\text{eff}}(t)} \|(\tilde{Q}_t)_{:j} - \tilde{S}_t\mathbf{w}\|_2^2 \ \text{ s.t. } J(\mathbf{w}) \leqslant J(W_{*j})$
```

3.1. The `GreedyLinCBwK` Algorithm

The algorithm has three key steps as illustrated in `GreedyStep` (Algorithm 2): parameter estimation, arm selection to receive reward and consumption vector, and dual update, with some steps picking the no-op arm.

Estimation. In each step, the algorithm estimates the reward vector $\widehat{\boldsymbol{\mu}}_t$ and constraint parameter $\widehat{W}_t$ based on a constrained least squares formulation as shown in `Estimate` (Algorithm 3). In the smoothed stochastic setting, one can show that the estimated parameters $\widehat{\boldsymbol{\mu}}_t, \widehat{W}_t$ approach the true parameters at a $O(1/\sqrt{T_{\text{eff}}(t)})$ rate using standard techniques. Showing a similar result for the smoothed adversarial settings needs more care, and our analysis leverages recent advances in such analysis (Sivakumar et al., 2020).

Arm Selection. The arm selection is based on

$$a_t = \operatorname*{argmax}_{a \in [K]} \ \mathbf{x}_t(a)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) \ ,$$

as long as $\mathbf{x}_t(a_t)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) \geqslant -2\zeta_t$ where

$$\zeta_t = O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)} \right) Z\beta \log K}{\sigma\sqrt{T_{\text{eff}}(t)}} \right) , \tag{10}$$

otherwise the no-op arm is selected. Such an arm choice was first explored in (Agrawal & Devanur, 2014b; 2016) who prescribed using $Z = \Omega(\text{OPT}/B)$ to obtain optimal regret rates, where OPT is the reward obtained by the optimal static policy. As we show in subsequent sections, the algorithms for the stochastic and adversarial settings differ in how OPT and by extension Z is estimated and used.

Dual Update. For a chosen arm a_t, the corresponding consumption vector $\mathbf{v}_t(a_t)$ needs to stay within the per step budget $\frac{B}{T}\mathbf{1}$, i.e., $\mathbf{v}_t(a_t) - \frac{B}{T}\mathbf{1} \leqslant \mathbf{0}$. Similar to (Agrawal & Devanur, 2016; Immorlica et al., 2019) we will assume the presence of a dummy time resource with B/T consumption in each round irrespective of the chosen arm. Following related work (Agrawal & Devanur, 2016; Abernethy et al., 2011), the dual update involves maximizing $\langle \boldsymbol{\theta}_{t+1}, \mathbf{v}_t(a_t) - \frac{B}{T}\mathbf{1} \rangle$ under the constraint $\boldsymbol{\theta}_{t+1} \in \Delta_{d+1}$ sequentially, which can be done with online mirror descent (OMD).

No-op Arm. Note that `GreedyStep` (Algorithm 2) chooses the no-op arm only when the objective for all arms is below $-2\zeta_t$ with ζ_t as in (10) which changes with time. The negative threshold $-2\zeta_t$ corresponds to the objective value estimation error when using $\widehat{\boldsymbol{\mu}}_t, \widehat{W}_t$ to estimate the objective instead of $\boldsymbol{\mu}_*, W_*$. In particular, in steps where `GreedyStep` chooses the no-op arm, an algorithm using $\boldsymbol{\mu}_*, W_*$ would also have chosen the no-op arm with high probability.

3.2. Primal-Dual Perspective

We develop an understanding of `GreedyLinCBwK` by considering the optimization problem in hindsight to find the optimum static policy over T steps. The purpose is to gain insights about key steps in the algorithm, and more rigorous technical results are presented in subsequent sections.

Consider the optimization problem:

$$\max_{\pi} \sum_{t=1}^{T} \boldsymbol{\mu}_*^\intercal X_t \pi(X_t) \quad \text{s.t.} \quad W_*^\intercal X_t \pi(X_t) - \frac{B}{T}\mathbf{1} \leqslant 0 \ , \ t \in [T] \tag{11}$$

Let $\mathbf{y}_t \in \mathbb{R}_+^{(d+1)\times 1}$ be the Lagrange mulipliers corresponding to the t-th constraint, and let $\boldsymbol{\theta}_t = \mathbf{y}_t / \|\mathbf{y}_t\|_1$ be the normalized Lagrange multipliers. The Lagrangian is given by:

$$\begin{aligned}\mathcal{L}(\pi, \boldsymbol{\theta}) &= \sum_{t=1}^{T} \boldsymbol{\mu}_*^\intercal X_t \pi(X_t) - \sum_{t=1}^{T} \left\langle \mathbf{y}_t, W_*^\intercal X_t \pi(X_t) - \frac{B}{T}\mathbf{1} \right\rangle \\ &= \sum_{t=1}^{T} \boldsymbol{\mu}_*^\intercal X_t \pi(X_t) - Z \sum_{t=1}^{T} \left\langle \boldsymbol{\theta}_t, W_*^\intercal X_t \pi(X_t) - \frac{B}{T}\mathbf{1} \right\rangle ,\end{aligned}$$

where Z is a suitable constant. Note that we are not showing the primal constraints $\pi(X_t) \in \Delta_{K+1}$ and dual constraints $\boldsymbol{\theta}_t \in \Delta_{d+1}$ explicitly to avoid clutter.

Scaling constant Z. Let OPT denote the (primal) optimal value for (11). Since the primal is a feasible bounded LP, strong duality holds (Bertsimas & Tsitsiklis, 1997; Boyd & Vandenberghe, 2005), and the solutions of the primal and dual problems match, i.e., is both OPT. With $\mathbf{y}_t^*$ denoting the optimal Lagrange dual parameters, by strong duality $\text{OPT} = \sum_{t=1}^{T} \langle \mathbf{y}_t^*, \frac{B}{T}\mathbf{1} \rangle$. Then, with $Z_t^* := \|\mathbf{y}_t^*\|_1$ and $\boldsymbol{\theta}_t^* = \mathbf{y}_t^* / \|\mathbf{y}_t^*\|_1$, we have

$$\begin{aligned}\text{OPT} &= \sum_{t=1}^{T} \langle \mathbf{y}_t^*, \frac{B}{T}\mathbf{1} \rangle = \sum_{t=1}^{T} Z_t^* \langle \boldsymbol{\theta}_t^*, \frac{B}{T}\mathbf{1} \rangle \\ &= \sum_{t=1}^{T} Z_t^* \frac{B}{T} = B \cdot Z^*,\end{aligned}$$

where $Z^* := \frac{1}{T}\sum_{t=1}^{T} Z_t^* = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{y}_t^*\|_1$, so that we have $Z^* = \frac{\text{OPT}}{B}$. Thus, the optimal scaling constant Z^* is simply the per step average of the sum of optimal Langrange multipliers $\|\mathbf{y}_t^*\|_1$, and satisfies $Z^* = \frac{\text{OPT}}{B}$ so that an estimate of OPT will yield an estimate of Z^*. In the online setting, $\mathbf{y}_t^*$ are not known, and are estimated sequentially. As a result, `GreedyLinCBwK` will keep an (running) estimate of OPT and hence Z based on (a) the current estimates $\widehat{\boldsymbol{\mu}}_t, \widehat{W}_t$ respectively of μ^*, W^*, and (b) the contexts $X_{1:t}$ observed till time step t.

Primal updates. The primal update would ideally be based on $\frac{\partial c L(\pi, \boldsymbol{\theta})}{\partial \pi(X_t)}$, i.e., gradient of the Lagrangian w.r.t. the policy. Since we can pull only one arm, i.e., a bandit step, the update will have to be based on partials w.r.t. $\pi(X_t)(a), a \in [K]$ and picking the arm with the largest gradient, i.e., a Frank-Wolfe, conditional gradient, or coordinate ascent type update (Frank & Wolfe, 1956). Note that the partial derivative is simply: $\frac{\partial \mathcal{L}(\pi, \boldsymbol{\theta})}{\partial \pi(X_t)(a)} = \mathbf{x}_t(a)^\intercal (\boldsymbol{\mu}_* - Z W_* \boldsymbol{\theta}_t)$, and the arm chosen is the one which maximizes this.

Dual updates. Since the true consumption vector at step t by pulling arm a is $\mathbf{v}_t(a)$ and since the primal is a maximization, the cumulative dual objective to be minimized is $-\sum_{t=1}^{T} \langle \boldsymbol{\theta}_t, \mathbf{v}_t(a) - \frac{B}{T}\mathbf{1} \rangle$ over $\boldsymbol{\theta}_t \in \Delta_{d+1}$. For convenience, we will equivalently consider maximizing $\sum_{t=1}^{T} \langle \boldsymbol{\theta}_t, \mathbf{v}_t(a) - \frac{B}{T}\mathbf{1} \rangle$. Since the problem needs to be solved sequentially, this becomes a simple example of OCO with a linear objective and simplex constraint , which can be solved by online mirror descent (OMD) or exponentiated gradient (EG) descent (Shalev-Shwartz et al., 2012) with sublinear regret.

Proposition 1 *With $g_t(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{v}_t(a) - \frac{B}{T}\mathbf{1} \rangle$, OMD with constraint $\boldsymbol{\theta}_t \in \Delta_{d+1}$ achieves the following regret:*

$$R_D(T) = \max_{\boldsymbol{\theta} \in \Delta_{d+1}} \sum_{t=1}^{T} g_t(\boldsymbol{\theta}) - \sum_{t=1}^{T} g_t(\boldsymbol{\theta}_t) = O\left(\sqrt{T \log(d)}\right) .$$

3.3. Main Result: Reward from `GreedyLinCBwK`

Theorem 1 lower bounds the reward obtained by the greedy algorithm. T_{stop} is either T or the first round any resource consumption exceeds B. The algorithm incurs additive regret $R(T)$ due to use of estimates $\widehat{\boldsymbol{\mu}}, \widehat{W}$ instead of $\boldsymbol{\mu}_*, W_*$ to compute the reward and constraint objectives. The reward can be lower bounded by either of two quantities. The primal update ensures the reward is greater than $\text{OPT}(T_{\text{stop}}) + Z(\gamma_a(T_{\text{stop}}) - \gamma_\pi(T_{\text{stop}}))$. Presence of the no-op arm ensures the reward is lower bounded by $Z\gamma_a(T_{\text{stop}})$.

Theorem 1 *Let $T_{stop} \leqslant T$ denote the stopping time of the algorithm. Let $\beta = \max_{t \in [T], a \in [K]} \|\mathbf{x}_t(a)\|_2$ be the maximum ℓ_2 norm of context vectors. Then with probability at least $1 - 2\delta$,*

$$\begin{aligned}\text{REW} \geqslant \max \big[\, &\text{OPT}(T_{stop}) + Z(\gamma_a(T_{stop}) - \gamma_\pi(T_{stop})), \\ &Z\gamma_a(T_{stop}) \big] - O(ZR(T)) \ ,\end{aligned}$$

where

$$\text{REW} = \sum_{t=1}^{T_{stop}} \langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) \rangle \tag{12}$$

$$\text{OPT}(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle \boldsymbol{\mu}_*, X_t \boldsymbol{\pi}^*(X_t) \rangle \tag{13}$$

$$\gamma_\pi(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle W_*^\intercal X_t \boldsymbol{\pi}^*(X_t), \boldsymbol{\theta}_t \rangle \tag{14}$$

$$\gamma_a(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle W_*^\intercal \mathbf{x}_t(a_t), \boldsymbol{\theta}_t \rangle \, . \tag{15}$$

and, with σ^2 being the Gaussian perturbation variance,

$$R(T) = O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right) . \tag{16}$$

4. Smoothed Stochastic LinCBwK

In this section, we discuss the algorithm for stochastic LinCBwK. The algorithm has two phases. In the initial warm start phase with T_0 time steps, the arms are chosen uniformly at random. The parameter Z is estimated at the end of the warm start phase. We further discuss estimation of Z in Section 4.1. In the exploit phase, greedy Algorithm 1 is run with a fixed Z computed at the end of the warm start phase and a budget $B' = B - \beta T_0$. This is due to Lemma 1 which bounds the maximum consumption for any resource per round to β. The execution of the algorithm stops either after T time steps or when one of the resource consumption exceeds B. Our algorithm differs from (Agrawal & Devanur, 2016) in two aspects. The algorithm chooses arms uniformly at random in the warm start phase instead of the more involved exploration technique employed in (Agrawal & Devanur, 2016). The algorithm computes greedy estimates of the reward and constraint objectives in each time step of the exploit phase instead of UCB estimates.

4.1. Estimation of Z

Recall the discussion in Section 3.2 where we establish $Z^* = \frac{\mathrm{OPT}_1}{B}$ where OPT_1 is the optimal static policy reward. To estimate Z after the warm start phase, we estimate the optimal reward $\widehat{\mathrm{OPT}}(T_0;\omega(T_0))$ where $\omega(T_0) = (X_t, r_t, \mathbf{v}_t, a_t)_{t=1}^{T_0}$ are observations in the warm start phase and extrapolate it. Let $\hat{\boldsymbol{\mu}}_{T_0}$ and $\widehat{W}_{T_0}$ denote the parameter estimates at the end of the warm start phase. Then,

$$\begin{aligned}\widehat{\mathrm{OPT}}(T_0;\omega(T_0)) &:= \max_{\pi} \sum_{t=1}^{T_0} \hat{\boldsymbol{\mu}}_{T_0}^{\intercal} X_t \boldsymbol{\pi}(X_t)\\ &\quad\text{s.t.} \quad \sum_{t=1}^{T_0} \widehat{W}_{T_0}^{\intercal} X_t \boldsymbol{\pi}(X_t) \leqslant B_0 + R(T_0)\ ,\\ \widehat{\mathrm{OPT}}(T;\omega(T_0)) &:= \frac{T}{T_0}\widehat{\mathrm{OPT}}(T_0)\ ,\end{aligned} \tag{17}$$

where $B_0 = (T_0/T)B$ and $R(T_0)$ is an upper bound on the regret $\sum_{t=1}^{T_0}(W_* - \widehat{W}_{T_0})X_t\pi(X_t)$ after the warm start phase. Note that $R(T_0) = O\left(\frac{\left(w(A)+\sqrt{\log(T_0 d/\delta)}\right)\beta\sqrt{T_0}\log K}{\sigma}\right)$ is the regret at the end of the warm start phase. We set $Z(T_0)$ as follows.

$$Z(T_0) = \frac{\left(\widehat{\mathrm{OPT}}(T;\omega(T_0)) + 2\left(\frac{T}{T_0}\right)R(T_0)\right)}{B} + 1\ .$$

Algorithm 4 Smoothed Stochastic LinCBwK

```
  Input: Budget $B \in \mathbb{R}_+$, Total timesteps $T$.
  Initialize $S_1 = [], y_1 = [], Q_1 = []$
  for $\tau = 1$ to $T_0$ do
    Select arm $\mathbf{x}_\tau(a_\tau)$ uniformly at random, observe noisy
    reward $r_\tau(a_\tau)$ and consumption vector $\mathbf{v}_\tau(a_\tau) \in$
    $\mathbb{R}^d$. Obtain $S_{\tau+1}, y_{\tau+1}, Q_{\tau+1}$ by appending $\mathbf{x}_\tau(a_\tau)$,
    $r_\tau(a_\tau)$, $\mathbf{v}_\tau(a_\tau)$ to $S_\tau, y_\tau, Q_\tau$.
  end for
  Let $B' = B - \beta T_0$. Compute parameter $Z$ such that
  $\frac{\mathrm{OPT}_1}{B'} \leqslant Z \leqslant O\left(\frac{\mathrm{OPT}_1}{B'} + 1\right)$.
  Run Algorithm GreedyLinCBwK with budget $B'$ from
  time $T_0 + 1$ to $T$.
```

We revisit the following result from (Agrawal & Devanur, 2016) on error bounds on estimator $\widehat{\mathrm{OPT}}(T;\omega(T_0))$ with respect to the optimal reward OPT_1. The bounds account for both the extrapolation error and use of $\hat{\boldsymbol{\mu}}_{T_0}, \widehat{W}_{T_0}$ instead of the true parameters for estimation.

Lemma 4 *[Lemma 5 in (Agrawal & Devanur, 2016)] For $R(T_0) = O\left(\frac{(w(A)+\sqrt{\log(T_0 d/\delta)})\beta\sqrt{T_0}\log K}{\sigma}\right)$, with probability at least $1-\delta$*

$$\begin{aligned}\widehat{\mathrm{OPT}}(T;\omega(T_0)) &\geqslant \mathrm{OPT}_1 - 2\left(\frac{T}{T_0}\right)R(T_0)\\ \widehat{\mathrm{OPT}}(T;\omega(T_0)) &\leqslant \mathrm{OPT}_1 + 9\left(\frac{T}{T_0}\right)R(T_0)\left(\frac{\mathrm{OPT}_1}{B}+1\right)\end{aligned} \tag{18}$$

Set $Z(T_0) = \frac{\left(\widehat{\mathrm{OPT}}(T;\omega(T_0))+2\left(\frac{T}{T_0}\right)R(T_0)\right)}{B} + 1$. Then with probability $1 - O(\delta)$,

$$\begin{aligned}Z(T_0) &\geqslant \frac{\mathrm{OPT}_1}{B} + 1\\ Z(T_0) &\leqslant \left(1 + \frac{11\left(\frac{T}{T_0}\right)R(T_0)}{B}\right)\left(\frac{\mathrm{OPT}_1}{B}+1\right)\ .\end{aligned} \tag{19}$$

Moreover if $B \geqslant \tilde{\Omega}\left(\frac{T}{\sqrt{T_0}}\frac{w(A)\log K}{\sigma}\right)$ then $Z(T_0) \leqslant O\left(\frac{\mathrm{OPT}_1}{B}+1\right)$.

4.2. Regret Analysis

Let $T_{\text{stop}} \leqslant T$ denote the stopping time step of the algorithm which is either T or the first time any of the resource consumption exceeds B. The final regret is obtained by combining the results of Lemma 4 and Theorem 1. Since arms are chosen at random we accrue linear regret in the warm start phase. Here $R_D(T)$ is the regret of the dual OCO algorithm from Proposition 1.

Theorem 2 *Assume* $B \geqslant \tilde{\Omega}\left(\max\left(T_0, \frac{T}{\sqrt{T_0}}\frac{w(A)\log K}{\sigma}\right)\right)$. *Then if* $Z(T_0)$ *is set as outlined in Lemma 4, we have*

$$\frac{\mathrm{OPT}_1}{B} + 1 \leqslant Z(T_0) \leqslant O\left(\frac{\mathrm{OPT}_1}{B} + 1\right) . \tag{20}$$

Moreover with probability at least $1 - 6\delta$*, following is the regret for the greedy Algorithm 4,*

$$\mathrm{regret}(T) \leqslant O\left(\left(\frac{\mathrm{OPT}_1}{B} + 1\right)\max\left(T_0, R(T)\right)\right) , \tag{21}$$

where $R(T) = O\left(\frac{\left(w(A)+\sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right)$.

We wrap up with special cases of our results.

Corollary 1 *Assume the variance of the Gaussian noise in the smoothed setting* $\sigma^2 = \Omega(1/m)$. *Then with probability at least* $1 - O(\delta)$,

- *If* $J(\cdot)$ *is the* ℓ_2^2 *norm,* $T_0 = \tilde{O}\left(m^{2/3}\sqrt{T}\right)$ *and* $B \geqslant \tilde{\Omega}(m^{2/3}T^{3/4})$,

$$\mathrm{regret}(T) \leqslant \tilde{O}\left(\left(\frac{\mathrm{OPT}_1}{B} + 1\right) m\sqrt{T}\right) \tag{22}$$

- *If* $J(\cdot) = \|\cdot\|_1$, $\boldsymbol{\mu}_*$ *and columns of* W_* *are s-sparse,* $T_0 = \tilde{O}((m \cdot s\log m)^{1/3}\sqrt{T})$ *and* $B \geqslant \tilde{\Omega}(m \cdot s\log m)^{1/3}T^{3/4})$,

$$\mathrm{regret}(T) \leqslant \tilde{O}\left(\left(\frac{\mathrm{OPT}_1}{B} + 1\right)\sqrt{m}\sqrt{s\log m}\sqrt{T}\right) \tag{23}$$

5. Smoothed Adversarial LinCBwK

In this section, we consider adversarial LinCBwK where the base context vectors $\boldsymbol{\nu}_t(a), a \in [K]$ of the smoothed contexts $\mathbf{x}_t(a) = \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a)$ are chosen by an adaptive adversary. The basic ideas behind the algorithm remain the same as in stochastic LinCBwK except for the estimation of the tradeoff parameter Z which balances optimizing the reward and consumption. But unlike the stochastic setting it is not possible to estimate OPT_2, the reward from the optimal feasible static policy, without knowledge of the adversarially chosen contexts $\boldsymbol{\nu}_t(a)$ over all time steps.

5.1. Estimation of Z

We address this challenge by adapting a variant of the standard doubling trick in online learning (Shalev-Shwartz et al., 2012). At any stage, the algorithm has an estimate of the value of OPT and by extension Z. In each time step t, the realized cumulative reward is compared with the estimate for OPT and the algorithm updates the estimate for OPT when the realized reward is double that of the current estimate for OPT, and also updates the estimate for Z. A related doubling trick adaptation is used in (Immorlica et al., 2019) for the adversarial multi-armed bandits with knapsacks problem although the algorithmic details are different.

Similar to the the stochastic setting, Algorithm 5 has a warm start phase with T_0 rounds. But unlike the stochastic setting the warm start phase is only required to get good estimates of the reward and constraint parameters. The exploit phase in Algorithm 5 is further subdivided into multiple epochs. Let T_1 denote the start time step of an epoch. Let $\omega(T_1) = (X_t, r_t, \mathbf{v}_t, a_t)_{t=1}^{T_1}$ denote data observed until T_1. At the first step of the epoch, $\widehat{\mathrm{OPT}}(T_1; \omega(T_1))$ and $Z(T_1)$ are computed with data $\omega(T_1)$ as follows:

$$\begin{aligned}\widehat{\mathrm{OPT}}(T_1; \omega(T_1)) &:= \max_{\pi} \sum_{t=1}^{T_1} \widehat{\boldsymbol{\mu}}_{T_1}^{\intercal} X_t \pi(X_t) \\ &\text{s.t.} \sum_{t=1}^{T_1} \widehat{W}_{T_1}^{\intercal} X_t \pi(X_t) \leqslant B + R(T_1) , \\ Z(T_1) &:= \frac{\widehat{\mathrm{OPT}}(T_1; \omega(T_1))}{2dB} .\end{aligned} \tag{24}$$

The value of $\widehat{\mathrm{OPT}}(T_2; \omega(T_2))$ is monitored in each time step $T_2 > T_1$ of an epoch until $\widehat{\mathrm{OPT}}(T_2; \omega(T_2)) \geqslant 2\widehat{\mathrm{OPT}}(T_1; \omega(T_1))$ at which point the epoch is terminated and a new epoch is started. Each epoch is allocated a budget $B'/2\lceil\log T\rceil$. In each time step of an epoch `GreedyStep` (Algorithm 2) is run until a new epoch is started or when the allocated budget for any resource is exhausted. In case of budget exhaustion, an arm is chosen uniformly at random with a fixed probability until a new epoch starts.

5.2. Regret Analysis

It is evident that the Z estimates improve with progression of epochs. We therefore focus on the last completed epoch and show there is sufficient reward to be collected from this epoch. Moreover we show that the greedy algorithm receives minimum $\Omega(1/(d\log T))$ fraction of the reward available in the final completed epoch. OPT_2 denotes the reward of the optimal static policy from Lemma 3.

Theorem 3 *Assume* $B \geqslant \Omega(T^{3/4})$. *Then with probability at least* $1 - 6\delta$ *following is a lower bound on the reward obtained by the greedy algorithm*

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}_2}{16d\lceil\log T\rceil} - O\left(\left(\frac{\mathrm{OPT}_2}{B} + 1\right)R(T)\right) , \tag{25}$$

where $R(T) = O\left(\frac{\left(w(A)+\sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right)$.

Algorithm 5 Smoothed Adversarial LinCBwK

```
Input: Budget $B$, number of time steps $T$
Initialize $S_1, y_1, Q_1 = []$
for $t = 1$ to $T_0$ do
    Select arm $\mathbf{x}_t(a_t)$ uniformly at random, observe noise reward $r_t(a_t)$ and consumption vector $\mathbf{v}_t(a_t)$. Get $S_{t+1}, y_{t+1}, Q_{t+1}$ by appending $\mathbf{x}_t(a_t), r_t(a_t), \mathbf{v}_t(a_t)$ to $S_t, y_t, Q_t$.
end for
Let $B' = B - T_0$.
Initialize epoch $j = \lfloor \log_2 \widehat{\mathrm{OPT}}(T_0; \omega(T_0)) \rfloor$
for each epoch do
    Let $T_1$ be the initial time step of epoch. Estimate $\widehat{\mathrm{OPT}}(T_1; \omega(T_1)), Z(T_1)$ using (24)
    Compute budget $B_0 = \frac{B'}{2\lceil \log T \rceil}$
    Initialize $\boldsymbol{\theta}_{T_1}$ using uniform distribution.
    for each time step $T_2$ in epoch do
        Recompute $\widehat{\mathrm{OPT}}(T_2; \omega(T_2))$ using (24)
        if $\widehat{\mathrm{OPT}}(T_2; \omega(T_2)) \geqslant 2^{j+1}$ then
            Increment $j = j + 1$ and start a new epoch
        else
            if any budget consumed then
                Run Algorithm GreedyStep with $B = B_0$ and $Z = Z(T_1)$.
            else
                Play any arm uniformly at random with probability $T^{-1/4}$, play the no-op arm otherwise
            end if
        end if
    end for
end for
```

(Immorlica et al., 2019) obtain $O(d\lceil \log T \rceil)$ competitive ratio for adversarial multi-armed bandits (MAB) with knapsacks and show it to be optimal. We match the bound for the linear contextual bandit setting with smoothed contexts with better additive regret rates. Note that (Immorlica et al., 2019) considered *oblivious* adversaries where we consider *adaptive* adversaries in the smoothed setting. Further, the additive regret in (Immorlica et al., 2019) is $\tilde{O}\left(KT^{7/4}/B\right)$ so for regret bounds to be meaningful requires the condition $\frac{B}{K}\,\mathrm{OPT} \geqslant \tilde{\Omega}(T^{7/4})$. Our setup requires much milder assumptions. We wrap up with special cases of our result.

Corollary 2 *Assume the variance of the Gaussian noise in the smoothed setting* $\sigma^2 = \Omega(1/m)$. *Let* OPT *be the optimal reward and* $\kappa \leqslant O(d\lceil \log T \rceil)$. *Then with probability at least* $1 - O(\delta)$,

- *If* $J(\cdot) = \|\cdot\|_2$ *and* $B \geqslant \tilde{\Omega}(T^{3/4})$,

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}_2}{\kappa} - \tilde{O}\left(\left(\frac{\mathrm{OPT}_2}{B} + 1\right) m\sqrt{T}\log K\right)$$

- *If* $R(\cdot) = \|\cdot\|_1$, $\boldsymbol{\mu}_*$ *and columns of* W_* *are s-sparse and* $B \geqslant \tilde{\Omega}(T^{3/4})$,

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}_2}{\kappa} - \tilde{O}\left(\left(\frac{\mathrm{OPT}_2}{B} + 1\right) \sqrt{m}\sqrt{s\log m}\sqrt{T}\right)$$

6. Conclusion

We considered the linear contextual bandit with knapsacks (LinCBwK) problem in the smoothed setting where the contexts chosen by nature, stochastically or adversarially, is perturbed by Gaussian noise. Such smoothed analysis, especially in the adversarial setting, is a mechanism to soften the worst case analysis and prior work has illustrated that simpler algorithms often work when avoiding the worst case. Our results in this paper continue this tradition. As the first work which analyzes smoothed adversarial LinCBwK with adaptive adversaries, we show that a combination of a greedy strategy to choose arms combined with a doubling trick gives a suitable algorithm for the setting. Scope for future work includes consideration of more flexible models for the rewards as well as budgets, including dynamic budgets.

Acknowledgements. We would like to thank the reviewers for valuable comments and questions. The research was supported by NSF grants IIS 21-31335, OAC 21-30835, DBI 20-21898, and a C3.ai research award.

References

Abernethy, J., Bartlett, P. L., and Hazan, E. Blackwell approachability and no-regret learning are equivalent. In *Conference on Learning Theory*, pp. 27–46. PMLR, 2011.

Agarwal, S., Wang, Z., and Ye, Y. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014a.

Agrawal, S. and Devanur, N. R. 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.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Babaioff, M., Dughmi, S., Kleinberg, R. D., and Slivkins, A. Dynamic pricing with limited supply. In *ACM Transactions on Economics and Computation*, volume 3, 2015.

Badanidiyuru, A., Kleinberg, R., and Singer, Y. Dynamic pricing with limited supply. In *ACM Transactions on Economics and Computation*, 2012.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with Norm Regularization. In *Neural Information Processing Systems (NIPS)*, 2014.

Bastani, H., Bayati, M., and Khosravi, K. Mostly exploration-free algorithms for contextual bandits. *arXiv:1704.09011*, 2018.

Bertsimas, D. and Tsitsiklis, J. *Introduction to Linear Optimization*. Athena Scientific, 1997.

Besbes, O. and Zeevi, A. J. Blind network revenue management. *Operations Research*, 60(6):1537–1550, 2012.

Bietti, A., Agarwal, A., and Langford, J. Practical evaluation and optimization of contextual bandit algorithms. *CoRR arXiv:1802.04064*, 2018.

Bird, S., Barocas, S., Crawford, K., Diaz, F., and Wallach, H. Exploring or exploiting? social and ethical implications of automonous experimentation. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2005.

Buchbinder, N. and Naor, J. S. Online primal-dual algorithms for covering and packing. *Mathematics of Operations Research*, 34(2), 2009.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Combes, R., Jiang, C., and Srikant, R. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1): 245–257, 2015.

Devanur, N. R. and Hayes, T. P. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *ACM Conference on Electronic Commerce (ACM-EC)*, 2009.

Devanur, N. R., Jain, K., Sivan, B., and Wilkens, C. A. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *ACM Conference on Electronic Commerce (ACM-EC)*, 2011.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V. S., and Stein, C. Online stochastic packing applied to display ad allocation. In *European Symposium on Algorithms (ESA)*, 2010.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics*, 3, 1956.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219. IEEE, 2019.

Kannan, S., Morgenstern, J., Roth, A., Waggoner, B., and Wu, Z. S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *CoRR arXiv:1801.04323*, 2018.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge University Press (To appear), 2019.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *International World Wide Web Conference (WWW)*, 2010.

Mehta, A. Online matching and ad allocation. *Foundations and Trends in Machine Learning*, 8(4):265–368, 2013.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized online matching. *Journal of the ACM*, 54(5), 2007.

Molinaro, M. and Ravi, R. Geometry of online packing linear problems. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2012.

Raghavan, M., Slivkins, A., Vaughan, J. W., and Wu, Z. S. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory (COLT)*, pp. 1724–1738, 2018.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2012.

Singla, A. and Krause, A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *International World Wide Web Conference (WWW)*, 2013.

Sivakumar, V., Wu, Z. S., and Banerjee, A. Structured linear contextual bandits: A sharp and geometric smoothed analysis. *arXiv:2002.11332v1*, 2020.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12(1-2):1–286, 2021.

Talagrand, M. *The Generic Chaining*. Springer Monographs in Mathematics. Springer Berlin, 2005.

Talagrand, M. *Upper and Lower Bounds of Stochastic Processes*. Springer, 2014.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1): 267–288, 1996.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press (To appear), 2019.

Williamson, D. P. and Shmoys, D. B. *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

A. Concentration Inequalities

We restate the following Hoeffding-type result for dependent variables from (Sivakumar et al., 2020).

Lemma 5 (Lemma 11 from (Sivakumar et al., 2020)) *Let $\{Z_t\}$ be a sub-Gaussian martingale difference sequence (MDS) and let $z_{1:t}$ denote a realization of $Z_{1:t}$. Let $\{a_t\}$ be a sequence of random variables such that $a_t = f_t(z_{1:(t-1)})$ for some sequence function f_t with $|a_t| \leqslant \alpha_t$ a.s. for suitable constants $\alpha_1, \dots, \alpha_T$. Then, for any $\tau > 0$, we have*

$$P\left(\left|\sum_{t=1}^{T} a_t z_t\right| \geqslant \tau\right) \leqslant 2\exp\left\{-\frac{\tau^2}{4c\kappa^2 \sum_{t=1}^{T} \alpha_t^2}\right\} , \tag{26}$$

for absolute constants $c > 0$ and where κ is the ψ_2-norm of the conditional subGaussian random variables.

Corollary 3 *Let $z_{1:T}$ denote denote a realization of a sub-Gaussian martingale difference sequence such that $\mathbb{E}[Z_i|z_{1:i-1}] = 0$ and $\|Z_i\|_{\psi_2} \leqslant \kappa_1$. Then using $\tau = \sqrt{4c\kappa_1^2 T \log(1/\delta)} = O(\sqrt{T\log(1/\delta)})$,*

$$P\left(\left|\sum_{t=1}^{T} z_t\right| \geqslant \sqrt{4c\kappa_1^2 T \log(1/\delta)}\right) \leqslant \delta . \tag{27}$$

B. Benchmark and Reward, Consumption Bounds

Lemma 1: *Let $\beta = \underset{t\in[T], a\in[K]}{\operatorname{argmax}} \|\mathbf{x}_t(a)\|_2$. Then with probability atleast $1-\delta$,*

$$\beta \leqslant O\left(1 + \sigma\sqrt{m\log(TK/\delta)}\right) . \tag{28}$$

Proof: We have $\mathbf{x}_t(a) = \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a)$ where $\boldsymbol{\nu}_t(a) \in \mathbb{B}_2^m$ and $\mathbf{g}_t(a) \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{m\times m})$. Therefore we have the following,

$$\|\mathbf{x}_t(a)\|_2 \leqslant \|\boldsymbol{\nu}_t(a)\|_2 + \|\mathbf{g}_t(a)\|_2 \leqslant 1 + \|\mathbf{g}_t(a)\|_2 . \tag{29}$$

To obtain bounds on $\|\mathbf{g}_t(a)\|_2$, note that $\mathbb{E}[\|\mathbf{g}_t(a)\|_2] \leqslant \sqrt{\mathbb{E}[\|\mathbf{g}_t(a)\|_2^2]} = \sigma\sqrt{m}$. Moreover we have the following result.

Lemma 6 (Lemma 1 from Laurent and Massart (Laurent & Massart, 2000)) *Support $x \sim \chi_m^2$, i.e., $x = \sum_{i=1}^m g_i^2$ for $g_i \in \mathcal{N}(0,1)$ independently, then*

$$P\left[x \geqslant m + 2\sqrt{m\epsilon} + 2\epsilon\right] \leqslant \exp(-\epsilon) . \tag{30}$$

Therefore from the above lemma the following can be deduced,

$$P[\|\mathbf{g}_t(a)\|_2 \geqslant 5\sigma\sqrt{m\epsilon}] \leqslant \exp(-\epsilon) .$$

Taking the max over all time steps by union bound argument,

$$P\left[\max_{t\in[T], a\in[K]} \|\mathbf{g}_t(a)\|_2 \geqslant 5\sigma\sqrt{m\epsilon}\right] \leqslant TK\exp(-\epsilon) .$$

Now choosing $\epsilon = \log(TK/\delta)$ and by simple arithmatic manipulations and combining with (29) we get the stated result. ■

Lemma 2: *Let $\overline{\mathrm{OPT}}_1 := \mathbb{E}_{\phi(T)}[\sum_{t=1}^T \boldsymbol{\mu}_*^\intercal X_t p_t^*(\phi(T))]$ denote the value of the optimal feasible dynamic policy p^* which knows parameters μ_*, W_* and distribution $\mathcal{D}$. Here $p_t^*(\phi(T))$ denotes the distribution over arms in the t-th time step. Then, the optimal static policy π^* in* (8) *satisfies: $Tr(\pi^*) \geqslant \overline{\mathrm{OPT}}_1$ and $Tv(\pi^*) \leqslant B\mathbf{1}$.*

Proof: Given the dynamic policy p^*, construct the following static context dependent policy π^* as follows.

$$\pi^*(X) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\phi(T)}[p_t^*(\phi(T))|X] .$$

Now by definition,

$$Tr(\pi^*) = T\mathbb{E}_{X\sim\mathcal{D}}[\mathbf{r}^\intercal \pi^*(X;\mathcal{D})] = \mathbb{E}_{\phi(T)}[\sum_{t=1}^{T} \mathbf{r}^\intercal p_t^*(\phi(T))] = \overline{\mathrm{OPT}}_1$$

$$T\mathbf{v}(\pi^*) = T\mathbb{E}_{X\sim\mathcal{D}}[V_t \pi^*(X;\mathcal{D})] = \mathbb{E}_{\phi(T)}[\sum_{t=1}^{T} V_t p_t^*(\phi(T))] \leqslant B\mathbf{1}\ .$$

■

Lemma 3: *Let* $\overline{\mathrm{OPT}}_2 := \sum_{t=1}^{T} \boldsymbol{\mu}_*^\intercal X_t p_t^*(\phi(T))$ *and* $\mathrm{OPT} := \sum_{t=1}^{T} \boldsymbol{\mu}_*^\intercal X_t \pi^*(X_t;\phi(T))$ *denote the value of the optimal feasible dynamic policy and the optimal static policy which has knowledge of the parameters* $\boldsymbol{\mu}_*, W_*$. *Then* $\mathrm{OPT} \geqslant \overline{\mathrm{OPT}}_2$.

Proof: For each context set X observed in $\phi(T)$, construct the following static context dependent policy,

$$\pi^*(X) := \mathrm{avg}[p_t^*(\phi(T))|X]\ .$$

It therefore follows by definition that $\overline{\mathrm{OPT}}_2 := \mathrm{OPT}_2$ and $\sum_{t=1}^{T} W_*^\intercal X_t \pi^*(X_t;\phi(T)) = \sum_{t=1}^{T} W_*^\intercal X_t p_t^*(\phi(T)) \leqslant \mathbf{1}B$. ■

C. Reward and Constraint Parameter Estimation

We prove upper bounds on parameter estimation errors in each round for the reward and constraint parameters. To be concise we will directly reference results from (Sivakumar et al., 2020) wherever possible.

C.1. Covariance Matrix Minimum Eigenvalue

Theorem 4 *Consider any time step t in Algorithm 1 when the no-op arm is not chosen. Let $\mathbf{x}_t(a_t)$ be the context corresponding to the chosen arm. The following is a lower bound on the minimum eigenvalue of the covariance matrix,*

$$\lambda_{\min}\left(\mathbb{E}_{\mathbf{x}_t(a_t)}\left[\mathbf{x}_t(a_t)\mathbf{x}_t(a_t)^\intercal\right]\right) \geqslant c_1 \frac{\sigma^2}{\log K} \tag{31}$$

Proof: Let $\mathbf{x}_t(a_t) = \boldsymbol{\nu}_t(a_t) + \mathbf{g}_t(a_t)$ where $a_t := \underset{a\in[K]}{\mathrm{argmax}}\ \mathbf{x}_t(a)^T(\hat{\boldsymbol{\mu}}_t - Z\hat{W}_t\boldsymbol{\theta}_t)$. Set $\boldsymbol{\phi}_t = \hat{\boldsymbol{\mu}}_t - Z\hat{W}_t\boldsymbol{\theta}_t$. By definition,

$$\begin{aligned}
&\lambda_{\min}\left(\mathbb{E}\left[\mathbf{x}_t(a_t)\mathbf{x}_t(a_t)^\intercal \,\middle|\, \mathbf{x}_t(a_t) = \underset{\mathbf{x}_t(a)}{\mathrm{argmax}}\langle \mathbf{x}_t(a), \boldsymbol{\phi}_t\rangle\right]\right) \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\intercal \left(\mathbb{E}\left[\mathbf{x}_t(a_t)\mathbf{x}_t(a_t)^\intercal \,\middle|\, \mathbf{x}_t(a_t) = \underset{\mathbf{x}_t(a)}{\mathrm{argmax}}\langle \mathbf{x}_t(a), \boldsymbol{\phi}_t\rangle\right]\right)\mathbf{w} \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \left(\mathbb{E}\left[\mathbf{w}^\intercal\mathbf{x}_t(a_t)\mathbf{x}_t(a_t)^\intercal\mathbf{w} \,\middle|\, \mathbf{x}_t(a_t) = \underset{\mathbf{x}_t(a)}{\mathrm{argmax}}\langle \mathbf{x}_t(a), \boldsymbol{\phi}_t\rangle\right]\right) \\
&\geqslant \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left(\left[\langle \mathbf{w}, \boldsymbol{\nu}_t(a_t) + \mathbf{g}_t(a_t)\rangle \,\middle|\, \mathbf{x}_t(a_t) = \underset{\mathbf{x}_t(a)}{\mathrm{argmax}}\langle \mathbf{x}_t(a), \boldsymbol{\phi}_t\rangle\right]\right) \\
&\geqslant \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left(\left[\langle \mathbf{w}, \mathbf{g}_t(a_t)\rangle \,\middle|\, \mathbf{g}_t(a_t) = \underset{\mathbf{g}_t(a)}{\mathrm{argmax}}\langle \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a), \boldsymbol{\phi}_t\rangle\right]\right)
\end{aligned} \tag{32}$$

Let $Q \in \mathbb{R}^{m\times m}$ be an orthogonal matrix such that $Q\boldsymbol{\phi}_t = (\|\boldsymbol{\phi}_t\|_2, 0, \ldots, 0)$. Also let $(\mathbf{g}_t(a_1), \ldots, \mathbf{g}_t(a_K)) = (Q^\intercal\boldsymbol{\epsilon}_t(a_1), \ldots, Q^\intercal\boldsymbol{\epsilon}_t(a_K))$. Due to rotational invariance property of multivariate Gaussian distributions $\boldsymbol{\epsilon}_t(a_i) \sim$

$N(0, \sigma^2 \mathbb{I}_{m\times m})$. The r.h.s. in (32) evaluates to the following,

$$\begin{aligned}
&\min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left[\langle \mathbf{g}_t(a_t), \mathbf{w}\rangle \;\middle|\; \mathbf{g}_t(a_t) = \underset{\mathbf{g}_t(a)}{\mathrm{argmax}} \langle \boldsymbol{\nu}_t(a) + \mathbf{g}_t(a), \boldsymbol{\phi}_t\rangle \right] \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left[\langle Q\mathbf{g}_t(a_t), Q\mathbf{w}\rangle \;\middle|\; \mathbf{g}_t(a_t) = \underset{\mathbf{g}_t(a)}{\mathrm{argmax}} \langle Q\boldsymbol{\nu}_t(a) + Q\mathbf{g}_t(a), Q\boldsymbol{\phi}_t\rangle \right] \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left[\langle \boldsymbol{\epsilon}_t(a_t), \mathbf{w}\rangle \;\middle|\; \boldsymbol{\epsilon}_t(a_t) = \underset{\boldsymbol{\epsilon}_t(a)}{\mathrm{argmax}} \langle Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a), Q\boldsymbol{\phi}_t\rangle \right] \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathrm{Var}\left[\langle \boldsymbol{\epsilon}_t(a_t), \mathbf{w}\rangle \;\middle|\; \boldsymbol{\epsilon}_t(a_t) = \underset{\boldsymbol{\epsilon}_t(a)}{\mathrm{argmax}} (Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a))_1 \right] \\
&= \min_{\mathbf{w}:\|\mathbf{w}\|_2=1} \left(\mathbf{w}_1^2 \mathrm{Var}(\boldsymbol{\epsilon}_t(a_t)_1) \;\middle|\; \boldsymbol{\epsilon}_t(a_t) = \underset{\boldsymbol{\epsilon}_t(a)}{\mathrm{argmax}} (Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a))_1 \right) + \\
&\qquad \left(\sum_{j=1}^{m} \mathbf{w}_j^2 \mathrm{Var}(\boldsymbol{\epsilon}_t(a_t)_j) \;\middle|\; \boldsymbol{\epsilon}_t(a_t) = \underset{\boldsymbol{\epsilon}_t(a)}{\mathrm{argmax}} (Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a))_1 \right)
\end{aligned} \tag{33}$$

Note that in the above, $(Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a))_1$ denotes the first coordinate of the vector $Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a)$. The below lemma establishes lower bounds on

$$\left(\mathrm{Var}(\boldsymbol{\epsilon}_t(a_t)_1) \;\middle|\; \boldsymbol{\epsilon}_t(a_t) = \underset{\boldsymbol{\epsilon}_t(a)}{\mathrm{argmax}} (Q\boldsymbol{\nu}_t(a) + \boldsymbol{\epsilon}_t(a))_1 \right)$$

Lemma 7 *Let $g_1, \ldots, g_K$ be independent Gaussian random variables sampled from $N(0, \sigma^2)$. Let $b_1, \ldots, b_K$ denote K random real numbers. Then,*

$$\left(Var(g_i) \;\middle|\; g_i = \underset{g_j}{\mathrm{argmax}}\ (g_j + b_j)\right) \geqslant \left(Var(g_i) \;\middle|\; g_i = \underset{g_j}{\mathrm{argmax}}\ g_j\right) \geqslant c_1 \frac{\sigma^2}{\log K}\ , \tag{34}$$

for some positive constant c_1.

We first prove the inequality $(\mathrm{Var}(g_i) \mid g_i = \underset{g_j}{\mathrm{argmax}}\ (g_j + b_j)) \geqslant (\mathrm{Var}(g_i) \mid g_i = \underset{g_j}{\mathrm{argmax}}\ g_j)$. Let $g_{(1)}, \cdots, g_{(K)}$ denote the order statistics of the Gaussian random variables such that $g_{(1)} \geqslant \cdots \geqslant g_{(K)}$. Now there are $K!$ permutations of the sum $b_i + g_{(j)},\ 1 \leqslant i, j \leqslant K$. Assume we have K buckets $A_1, \cdots, A_K$ and we partition the $K!$ permutations into the K buckets such that $A_1 \cup \cdots \cup A_K$ contains all the $K!$ permutations and $A_k \cap A_{k'} = \phi,\ 1 \leqslant k, k' \leqslant K, k \neq k'$.

Consider a single permutation when $b_i + g_{(j)}$ has the highest value for some $1 \leqslant i, j \leqslant K$. Assume ties are broken randomly with equal probability. For example, if $b_i + g_{(j)}, b_{i'} + g_{(j')}$, have equal values either is selected at random with equal probability.

Case 1: If $j = K$ we assign the permutation to the bucket A_K.

Case 2: If case 1 is not satisfied, consider indices (i', i'', j') such that the following conditions are satisfied:

1. $1 \leqslant i', i'', j' < K,\ i' \neq i'',\ j \leqslant j'$
2. $b_i + g_{(j)} \geqslant b_{i'} + g_{(j')}$ and $b_i + g_{(j)} > b_{i''} + g_{(j'+1)}$ in the current permutation
3. $b_i + g_{(j')} > b_{i'} + g_{(j)}$ in the permutation derived from the current permutation by swapping $g_{(j)}, g_{(j')}$ but $b_i + g_{(j'+1)} < b_{i''} + g_{(j)}$ in the permutation obtained by swapping $g_{(j)}, g_{(j'+1)}$

Condition 2, basically, is the assumption that $b_i + g_{(j)}$ has the highest value for the current permutation. Condition 3 finds an index $j' > j$ such in spite of swapping $g_{(j')}$ and $g_{(j)}$ to obtain a new permutation, $b_i + g_{(j')}$ has the highest value, but if

$g_{(j'+1)}$ and $g_{(j)}$ are swapped $b_i + g_{(j'+1)}$ no longer has the highest value. Note that one possible assignment is $i' = i$ and $j' = j$

In this case we assign the permutation to bucket $A_{j'}$.

Now by construction, the buckets partition the $K!$ permutations, i.e., $A_k \cap A_{k'} = \phi,\ k \neq k'$. Also in bucket A_j, $g_{(j')}$'s are present in equal proportion $\forall\ j' \leqslant j$. This is because of the following reason. Bucket A_j has no permutation with $b_{i'} + g_{(j')},\ j' > j$ with highest value due to the construction above. Let $b_i + g_{(j)}$ have the highest value for a permutation from bucket j. Then since $g_{(j')} > g_{(j)},\ \forall j' < j$, the permutation obtained by swapping $g_{(j)}, g_{(j')}$ will have $b_i + g_{(j')}$ as the highest value and all the permutations are equally probable.

Therefore,

$$(\mathrm{Var}(g_i) \mid g_i = \operatorname*{argmax}_{g_j}\,(g_j + b_j)) = \sum_{i=1}^{K} (\mathrm{Var}(g_i) \mid g_i \geqslant g_{(j)}) P(A_j)\ , \tag{35}$$

where $P(A_j)$ is the proportion of the $K!$ permutations in bucket j. Finally, we observe that $(Var(g_i) \mid g_i \geqslant g_{(j)})$ has the lowest value when $j = 1$ and hence $\sum_{i=1}^{K} (\mathrm{Var}(g_i) \mid g_i \geqslant g_{(j)}) P(A_j)$ has the lowest value when $P(A_1) = 1$ and $P(A_j) = 0,\ j \neq 1$, thus proving

$$(\mathrm{Var}(g_i) \mid g_i = \operatorname*{argmax}_{g_j}\,(g_j + b_j)) \geqslant (\mathrm{Var}(g_i) \mid g_i = \operatorname*{argmax}_{g_j} g_j)\ . \tag{36}$$

■

C.2. Reward and Constraint Parameters Estimation Error Bounds

The following result is borrowed from (Sivakumar et al., 2020). Note that the result in (Sivakumar et al., 2020) was proved for the setting when the base vectors are chosen adversarially and by extension are also applicable to the simpler setting when the base vectors are chosen stochastically.

Theorem 5 ***[Theorem 7 in (Sivakumar et al., 2020)]*** *Consider any time step t after the warm start phase. Let $T_{\mathit{eff}}(t)$ denote the number of time steps before t when the no-op arm was not chosen. Assume the rows of the design matrix in time step t satisfy the following covariance minimum eigenvalue condition.*

$$\lambda_{\min}\left(\mathbb{E}_{\mathbf{x}_t(a_t)}\left[\mathbf{x}_t(a_t)\mathbf{x}_t(a_t)^\intercal\right]\right) \geqslant c_1 \frac{\sigma^2}{\log K}\ . \tag{37}$$

Then with probability atleast $1 - \delta/T$ following is the estimation error bounds for the reward and constraint parameters.

$$\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_*\|_2 \leqslant O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\log K}{\sigma\sqrt{T_{\mathit{eff}}(t)}}\right) \tag{38}$$

$$\max_{1 \leqslant j \leqslant d} \|(W_t)_{:j} - (W_*)_{:j}\|_2 \leqslant O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\log K}{\sigma\sqrt{T_{\mathit{eff}}(t)}}\right)\ . \tag{39}$$

Proof: Theorem 4 proves the covariance minimum eigenvalue condition. The parameter estimation error upper bounds can be obtained using arguments from Theorem 7 in (Sivakumar et al., 2020) with slight modifications. ■

D. Greedy Algorithm Subroutine Analysis

We provide proof for Theorem 1.

Theorem 1 *Let $T_{stop} \leqslant T$ denote the stopping time of the algorithm. Let $\beta = \max_{t\in[T],a\in[K]} \|\mathbf{x}_t(a)\|_2$ be the maximum ℓ_2 norm of context vectors. Then with probability at least $1-2\delta$,*

$$\begin{aligned}\mathrm{REW} \geqslant \max\big[\, &\mathrm{OPT}(T_{stop}) + Z(\gamma_a(T_{stop}) - \gamma_\pi(T_{stop})),\\ &Z\gamma_a(T_{stop})\big] - O(ZR(T))\ ,\end{aligned}$$

where

$$\mathrm{REW} = \sum_{t=1}^{T_{stop}} \langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t)\rangle \tag{12}$$

$$\mathrm{OPT}(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle \boldsymbol{\mu}_*, X_t\boldsymbol{\pi}^*(X_t)\rangle \tag{13}$$

$$\gamma_\pi(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle W_*^\intercal X_t\boldsymbol{\pi}^*(X_t), \boldsymbol{\theta}_t\rangle \tag{14}$$

$$\gamma_a(T_{stop}) = \sum_{t=1}^{T_{stop}} \langle W_*^\intercal \mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle\ . \tag{15}$$

and, with σ^2 being the Gaussian perturbation variance,

$$R(T) = O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right)\ . \tag{16}$$

Proof: We first revise the notations that will be used in the proof. Let $T_{\text{eff}} = T_{\text{eff}}(T_{\text{stop}})$ be the effective number of time steps where any arm other than no-op is chosen after T time steps. The arm chosen by the algorithm at time step t satisfies one of the following two conditions.

1. The no-op arm is chosen by the algorithm when $\mathbf{x}_t(a'_t)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) < -2\zeta_t$ with $a'_t = \underset{a\in[K]}{\operatorname{argmax}}\ x_t(a)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t)$,
2. Arm a_t is chosen when $\mathbf{x}_t(a_t)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) \geqslant -2\zeta_t$,

where $\zeta_t = O\left(\frac{\left(w(A)+\sqrt{\log(Td/\delta)}\right)Z\beta\log K}{\sigma\sqrt{T_{\text{eff}}(t)}}\right)$.

Let Υ_{no} denote the set of time steps when condition 1 is triggered and Υ_a denotes the set of time steps when condition 2 is triggered. The size of the sets $|\Upsilon_{\text{no}}| = T_{\text{stop}} - T_{\text{eff}}$ and $|\Upsilon_a| = T_{\text{eff}}$.

Assume Algorithm 1 has knowledge of $\boldsymbol{\mu}_*, W_*$. Then in time step t arm $\mathbf{x}_t(a_t^*)$ will be chosen and both of the below conditions are true,

$$\begin{aligned}&\sum_{t=1}^{T_{\text{stop}}} \left(\langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t^*)\rangle - Z\langle W_*^\intercal \mathbf{x}_t(a_t^*), \boldsymbol{\theta}_t\rangle\right) \geqslant 0\\ &\sum_{t=1}^{T_{\text{stop}}} \left(\langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t^*)\rangle - Z\langle W_*^\intercal \mathbf{x}_t(a_t^*), \boldsymbol{\theta}_t\rangle\right) \geqslant \sum_{t=1}^{T_{\text{stop}}} \left(\langle \boldsymbol{\mu}_*, X_t\pi^*(X_t)\rangle - Z\langle W_*^\intercal X_t\pi^*(X_t), \boldsymbol{\theta}_t\rangle\right)\ .\end{aligned} \tag{40}$$

The first line is because we have the no-op option which has 0 rewards and $\mathbf{0}$ consumption. Note that since we have knowledge of $\boldsymbol{\mu}_*, W_*$ we no longer need to consider the negative threshold ζ_t which was to account for parameter estimation errors.

The below lemma shows that with high probability when the algorithm chooses the no-op arm in rounds in Υ_{no} the actual arm chosen with foreknowledge of $\boldsymbol{\mu}_*, W_*$ will also be the no-op arm.

Lemma 8 *Denote optimal arm in each round by a_t^* which is the arm that would have been chosen if the true parameters $\boldsymbol{\mu}_*, W_*$ were known and optimal arm other than no-op in each time step by $a_t^{'*} := \operatorname*{argmax}_{a\in[K]}\ \mathbf{x}_t(a)^\intercal(\boldsymbol{\mu}_* - ZW_*\boldsymbol{\theta}_t)$. Let $a_t' := \operatorname*{argmax}_{a\in[K]} \mathbf{x}_t(a)^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t)$ be the optimal arm without considering no-op chosen in the Algorithm. Then for any round t in Υ_{no} with $\zeta_t = O\left(\frac{\left(w(A)+\sqrt{\log(Td/\delta)}\right)Z\beta\log K}{\sigma\sqrt{T_{\mathit{eff}}(t)}}\right)$, we have*

$$\mathbf{x}_t(a_t')^\intercal(\widehat{\boldsymbol{\mu}}_t - Z\widehat{W}_t\boldsymbol{\theta}_t) < -2\zeta_t\ . \tag{41}$$

Then with probability atleast $1-\frac{2\delta}{T}$, a_t^* is also the no-op arm.

Proof: For the algorithm with knowledge of $\boldsymbol{\mu}_*, W_*$ to choose the no-op arm we have to prove $\langle\boldsymbol{\mu}_*, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle W_*^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle < 0$.

$$\begin{aligned}
\langle\boldsymbol{\mu}_*, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle W_*^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle &= \langle\boldsymbol{\mu}_* - \widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle(W_* - \widehat{W}_t)^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle \\
&\quad + \langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle \\
&< \langle\boldsymbol{\mu}_* - \widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle(W_* - \widehat{W}_t)^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle \\
&\quad + \underbrace{\langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t')\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t'), \boldsymbol{\theta}_t\rangle}_{-2\zeta_t} \\
&< \underbrace{\|\mathbf{x}_t(a_t^{'*})\|_2\|\boldsymbol{\mu}_* - \widehat{\boldsymbol{\mu}}_t\|_2}_{<\zeta_t} + Z\underbrace{\|\mathbf{x}_t(a_t^{'*})\|_2\|W_* - \widehat{W}_t\|_\infty\|\boldsymbol{\theta}_t\|_1}_{<\zeta_t} - 2\zeta_t \\
&< 0 \quad \text{with probability} \quad 1 - \frac{2\delta}{T}
\end{aligned}$$

In the second line we use $\langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t^{'*})\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t^{'*}), \boldsymbol{\theta}_t\rangle \leqslant \langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t')\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t'), \boldsymbol{\theta}_t\rangle$ by definition. Also since the no-op arm was chosen we have $\langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t')\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t'), \boldsymbol{\theta}_t\rangle < -2\zeta_t$. In the third line we use $\|\mathbf{x}_t(a_t^*)\|_2 \leqslant \beta$ from Lemma 1 and use estimation error bounds from Theorem 5. ■

Now it can be deduced from Lemma 8 using a union bound argument over rounds in Υ_{no} when the no-op arm is chosen with high probability the optimal arm is also the no-op arm. Therefore with probability $1 - \frac{2\delta(T_{\text{stop}} - T_{\text{eff}})}{T}$,

$$\sum_{t\in\Upsilon_{\text{no}}}\langle\boldsymbol{\mu}_*, \mathbf{x}(a_t)\rangle = \sum_{t\in\Upsilon_{\text{no}}}\langle\boldsymbol{\mu}_*, \mathbf{x}(a_t^*)\rangle = \sum_{t\in\Upsilon_{\text{no}}}\langle W_*^\intercal\mathbf{x}(a_t), \boldsymbol{\theta}_t\rangle = \sum_{t\in\Upsilon_{\text{no}}}\langle W_*^\intercal\mathbf{x}(a_t^*), \boldsymbol{\theta}_t\rangle = 0\ . \tag{42}$$

The following is true for time steps in set Υ_a due to the greedy choice made in Algorithm 2.

$$\sum_{t\in\Upsilon_a}\left(\langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t)\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle\right) \geqslant \sum_{t\in\Upsilon_a}\left(\langle\widehat{\boldsymbol{\mu}}_t, \mathbf{x}_t(a_t^*)\rangle - Z\langle\widehat{W}_t^\intercal\mathbf{x}_t(a_t^*), \boldsymbol{\theta}_t\rangle\right)$$

$$\Rightarrow \sum_{t\in\Upsilon_a}\left(\langle\boldsymbol{\mu}_*, \mathbf{x}_t(a_t)\rangle - Z\langle W_*^\intercal\mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle\right) + \underbrace{\sum_{t\in\Upsilon_a}\langle\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)\rangle}_{\leqslant R_{\boldsymbol{\mu}}(T_{\text{stop}})} \tag{43}$$

$$- Z\underbrace{\sum_{t\in\Upsilon_a}\langle(\widehat{W}_t - W_*)^\intercal(\mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)), \boldsymbol{\theta}_t\rangle}_{\leqslant R_W(T_{\text{stop}})}$$

$$\geqslant \sum_{t\in\Upsilon_a}\left(\langle\boldsymbol{\mu}_*, \mathbf{x}_t(a_t^*)\rangle - Z\langle W_*^\intercal\mathbf{x}_t(a_t^*), \boldsymbol{\theta}_t\rangle\right) \tag{44}$$

Combining results (40), (42) and (44) leads to the following with probability atleast $1 - \frac{2\delta(T_{\text{stop}} - T_{\text{eff}})}{T}$.

$$\underbrace{\sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t)\rangle}_{\text{REW}} \geqslant Z\underbrace{\sum_{t=1}^{T_{\text{stop}}}\langle W_*^{\intercal}\mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle}_{\gamma_a(T_{\text{stop}})} - \underbrace{\sum_{t\in\Upsilon_a}\langle \widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)\rangle}_{\leqslant R_{\boldsymbol{\mu}}(T_{\text{stop}})} - Z\underbrace{\sum_{t\in\Upsilon_a}\langle (W_* - \widehat{W}_t)^{\intercal}(\mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)), \boldsymbol{\theta}_t\rangle}_{\leqslant R_W(T_{\text{stop}})}$$

$$\begin{aligned}\underbrace{\sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t)\rangle}_{\text{REW}} \geqslant{}& \underbrace{\sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, X_t\pi^*(X_t)\rangle}_{\text{OPT}(T_{\text{stop}})} - \underbrace{\sum_{t\in\Upsilon_a}\langle \widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)\rangle}_{\leqslant R_{\boldsymbol{\mu}}(T_{\text{stop}})} - Z\underbrace{\sum_{t\in\Upsilon_a}\langle (W_* - \widehat{W}_t)^{\intercal}(\mathbf{x}_t(a_t) - \mathbf{x}_t(a_t^*)), \boldsymbol{\theta}_t\rangle}_{\leqslant R_W(T_{\text{stop}})} \\ &+ Z\underbrace{\sum_{t=1}^{T}\langle W_*^{\intercal}\mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle}_{\gamma_a(T_{\text{stop}})} - Z\underbrace{\sum_{t=1}^{T}\langle W_*^{\intercal}X_t\pi^*(X_t), \boldsymbol{\theta}_t\rangle}_{\gamma_\pi(T_{\text{stop}})}\ .\end{aligned}\tag{45}$$

The below lemma upper bounds the regret $R_{\boldsymbol{\mu}}(T_{\text{stop}}), R_W(T_{\text{stop}})$.

Lemma 9 *Let,*

$$R(T) = O\left(\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right),\tag{46}$$

where $\beta = \max_{t\in[T],a\in[K]}\|\mathbf{x}_t(a)\|_2 \leqslant O\left(1 + \sigma\sqrt{m\log(TK/\delta)}\right)$. *With probability atleast* $1 - \frac{2\delta T_{\mathit{eff}}}{T}$,

$$R_{\boldsymbol{\mu}}(T_{\mathit{stop}}) \leqslant R(T) \qquad R_W(T_{\mathit{stop}}) \leqslant R(T)\ .\tag{47}$$

Proof: Let Υ_a denote the set of time steps when any arm other than no-op was chosen. Then we have the following,

$$\begin{aligned}R_{\boldsymbol{\mu}}(T_{\text{stop}}) &\leqslant 2\beta\sum_{t\in\Upsilon_a}\|\boldsymbol{\mu}_* - \widehat{\boldsymbol{\mu}}_t\|_2 \\ &\leqslant 2\beta\sum_{t\in\Upsilon_a} c\cdot\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\log K}{\sigma\sqrt{T_{\text{eff}}(t)}} \quad \text{....... with prob. } \geqslant 1 - \frac{\delta T_{\text{eff}}}{T}\text{, using Theorem 5 and union bound} \\ &\leqslant 2\beta c\cdot\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\sqrt{T_{\text{eff}}(t)}\log K}{\sigma} \leqslant R(T)\ ,\end{aligned}\tag{48}$$

where c is a positive constant and in line 2 we use the reward estimation error rates from Theorem 5 together with a union bounding argument.

Similarly for bounding the constraint regret consider the following,

$$\begin{aligned}R_W(T_{\text{stop}}) &\leqslant 2\beta\sum_{t\in\Upsilon_a}\|\widehat{W}_t - W_*\|_\infty\|\boldsymbol{\theta}\|_1 \\ &\leqslant 2\beta\sum_{t\in\Upsilon_a}\max_{1\leqslant j\leqslant d}\|(\widehat{W}_t)_{:j} - (W_*)_{:j}\|_2 \quad \text{.... since } \|\boldsymbol{\theta}\|_1 \leqslant 1 \\ &\leqslant 2\beta\sum_{t\in P_2} c\cdot\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\log K}{\sigma\sqrt{T_{\text{eff}}(t)}} \quad \text{....... with prob. } \geqslant 1 - \frac{\delta T_{\text{eff}}}{T}\text{, Theorem 5 and union bound} \\ &\leqslant 2\beta c\cdot\frac{\left(w(A) + \sqrt{\log(Td/\delta)}\right)\sqrt{T_{\text{eff}}(t)}\log K}{\sigma} \leqslant R(T)\ .\end{aligned}\tag{49}$$

Now from equation (45) and the result of Lemma 9 we get the advertised result with the observation that $T_{\text{stop}} \leqslant T$. ■

E. Regret For Smoothed Stochastic Linear Bandits with Knapsacks

Lemma 4 *[Lemma 5 in (Agrawal & Devanur, 2016)] For $R(T_0) = O\left(\frac{(w(A)+\sqrt{\log(T_0 d/\delta)})\beta\sqrt{T_0}\log K}{\sigma}\right)$, with probability at least $1-\delta$*

$$\begin{aligned} \widehat{\mathrm{OPT}}(T;\omega(T_0)) &\geqslant \mathrm{OPT}_1 - 2\left(\frac{T}{T_0}\right)R(T_0) \\ \widehat{\mathrm{OPT}}(T;\omega(T_0)) &\leqslant \mathrm{OPT}_1 + 9\left(\frac{T}{T_0}\right)R(T_0)\left(\frac{\mathrm{OPT}_1}{B}+1\right) \end{aligned} \tag{18}$$

Set $Z(T_0) = \frac{\left(\widehat{\mathrm{OPT}}(T;\omega(T_0))+2\left(\frac{T}{T_0}\right)R(T_0)\right)}{B} + 1$. Then with probability $1-O(\delta)$,

$$\begin{aligned} Z(T_0) &\geqslant \frac{\mathrm{OPT}_1}{B} + 1 \\ Z(T_0) &\leqslant \left(1 + \frac{11\left(\frac{T}{T_0}\right)R(T_0)}{B}\right)\left(\frac{\mathrm{OPT}_1}{B}+1\right) . \end{aligned} \tag{19}$$

Moreover if $B \geqslant \tilde{\Omega}\left(\frac{T}{\sqrt{T_0}}\frac{w(A)\log K}{\sigma}\right)$ then $Z(T_0) \leqslant O\left(\frac{\mathrm{OPT}_1}{B}+1\right)$.

Proof: We borrow work from Lemma 5 in (Agrawal & Devanur, 2016). The only difference compared to (Agrawal & Devanur, 2016) is the step to bound the empirical reward and consumptions with the expectation since we have Gaussian perturbations which are no longer bounded. We have to prove the following.

$$\begin{aligned} \left|\sum_{t=1}^{T_0}\langle \boldsymbol{\mu}_*, X_t\pi(X_t)\rangle - \mathbb{E}_X[\langle \boldsymbol{\mu}_*, X_t\pi(X_t)\rangle]\right| &\leqslant R(T_0) \\ \max_{1\leqslant j\leqslant d}\left|\sum_{t=1}^{T_0}\langle (W_*)_{:j}, X_t\pi(X_t)\rangle - \mathbb{E}_X[\langle (W_*)_{:j}, X_t\pi(X_t)\rangle]\right| &\leqslant R(T_0) \end{aligned} \tag{50}$$

This is equivalent to proving bounds on $\left|\sum_{t=1}^{T_0}\langle \boldsymbol{\mu}_*, X_t\pi(X_t) - \mathbb{E}_X[X_t\pi(X_t)]\rangle\right| = |\sum_{t=1}^{T_0} z_t|$ where $z_t = \langle \boldsymbol{\mu}_*, X_t\pi(X_t) - \mathbb{E}_X[X_t\pi(X_t)]\rangle$. Note that z_t are (conditionally on history) sub-Gaussian with sub-Gaussian norm $c_2(1+\sigma) = O(1)$. This is because the base vectors of the contexts are bounded in $\mathbb{B}_2^m$ and σ^2 is the variance of the Gaussian perturbations. Therefore applying Corollary 3 we get,

$$P\left(\left|\sum_{t=1}^{T_0}\langle \boldsymbol{\mu}_*, X_t\pi(X_t) - \mathbb{E}_X[X_t\pi(X_t)]\rangle\right| \geqslant \sqrt{4c_2^2(1+\sigma)^2\log(T_0 d/\delta)}\right) \leqslant (\delta/T_0 d) . \tag{51}$$

Similarly,

$$P\left(\left|\sum_{t=1}^{T_0}\langle (W_*)_{:j}, X_t\pi(X_t) - \mathbb{E}_X[X_t\pi(X_t)]\rangle\right| \geqslant \sqrt{4c_2^2(1+\sigma)^2\log(T_0 d/\delta)}\right) \leqslant (\delta/T_0 d) . \tag{52}$$

Now taking a union bound over all $1 \leqslant j \leqslant d$ and with the observation $\sqrt{4c_2^2(1+\sigma)^2\log(T_0 d/\delta)} \leqslant R(T_0)$ we obtain 50 with probability atleast δ/T_0. The rest of the proof follows Lemma 5 in (Agrawal & Devanur, 2016) and Lemmas F.4,F.6 in (Agrawal & Devanur, 2014a). ■

We provide the proof for the final regret bounds in the stochastic setting for the greedy algorithm.

Theorem 2 *Assume $B \geqslant \tilde{\Omega}\left(\max\left(T_0, \frac{T}{\sqrt{T_0}}\frac{w(A)\log K}{\sigma}\right)\right)$. Then if $Z(T_0)$ is set as outlined in Lemma 4, we have*

$$\frac{\mathrm{OPT}_1}{B} + 1 \leqslant Z(T_0) \leqslant O\left(\frac{\mathrm{OPT}_1}{B}+1\right) . \tag{20}$$

Moreover with probability at least $1-6\delta$, following is the regret for the greedy Algorithm 4,

$$\mathrm{regret}(T) \leqslant O\left(\left(\frac{\mathrm{OPT}_1}{B}+1\right)\max\left(T_0, R(T)\right)\right) \ , \tag{21}$$

where $R(T) = O\left(\frac{\left(w(A)+\sqrt{\log(Td/\delta)}\right)\beta\sqrt{T}\log K}{\sigma}\right)$.

Proof: The condition on B and $Z(T_0)$ is a consequence of result of Lemma 4. We now focus on bounding the regret. We accrue maximum $O(\beta T_0)$ regret in the warm start phase. For characterizing regret during the exploit phase, let's start with the lower bound for REW from Theorem 1. With probability atleast $1-2\delta$,

$$\mathrm{REW} \geqslant \mathrm{OPT}(T_{\text{stop}}) + Z(T_0)\gamma_a(T_{\text{stop}}) - Z(T_0)\gamma_\pi(T_{\text{stop}}) - O(Z(T_0)\cdot R(T)) \ , \tag{53}$$

where

$$\mathrm{REW} = \sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t)\rangle \tag{54}$$

$$\mathrm{OPT}(T_{\text{stop}}) = \sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, X_t\boldsymbol{\pi}^*(X_t)\rangle \tag{55}$$

$$\gamma_\pi(T_{\text{stop}}) = \sum_{t=1}^{T_{\text{stop}}}\langle W_*^{\intercal}X_t\boldsymbol{\pi}^*(X_t), \boldsymbol{\theta}_t\rangle \tag{56}$$

$$\gamma_a(T_{\text{stop}}) = \sum_{t=1}^{T_{\text{stop}}}\langle W_*^{\intercal}\mathbf{x}_t(a_t), \boldsymbol{\theta}_t\rangle \ . \tag{57}$$

For stochastic LinCBwK $t=1$ references first round after warm start phase T_0+1, and if the algorithm stops at time step T' then $T_{\text{stop}} = T' - T_0$

Taking expectations w.r.t. X on both sides, with probability atleast $1-2\delta$

$$\mathbb{E}_X[\mathrm{REW}] \geqslant \mathbb{E}_X[\mathrm{OPT}(T_{\text{stop}})] + Z(T_0)\mathbb{E}_X[\gamma_a(T_{\text{stop}}) - \gamma_\pi(T_{\text{stop}})] - O(Z(T_0)\cdot R(T)) \ . \tag{58}$$

Also using definition of optimal regret,

$$\mathbb{E}_X[\mathrm{OPT}(T_{\text{stop}})] = \mathbb{E}_{X_t}\left[\sum_{t=1}^{T_{\text{stop}}}\langle \boldsymbol{\mu}_*, X_t\pi^*(X_t)\rangle\right] = \frac{T_{\text{stop}}}{T}\,\mathrm{OPT}_1 \ . \tag{59}$$

Lemma 10 *Following is upper bound on $\mathbb{E}_X[\gamma_a(T_{stop}) - \gamma_\pi(T_{stop})]$,*

$$\mathbb{E}_X[\gamma_a(T_{stop}) - \gamma_\pi(T_{stop})] \geqslant B\left(1-\frac{T_{stop}}{T}\right) - \beta T_0 - R_D(T_{stop}) \ , \tag{60}$$

where $R_D(T_{stop})$ is the regret of the OCO algorithm.

Proof: We make the following observations.

$$\begin{aligned}
\mathbb{E}_X[\gamma_a(T_{\text{stop}}) - \gamma_\pi(T_{\text{stop}})] &= \mathbb{E}_X[\sum_{t=1}^{T_{\text{stop}}} \langle W_*^\intercal(\mathbf{x}_t(a_t) - X_t\boldsymbol{\pi}^*(X_t)), \boldsymbol{\theta}_t\rangle] \qquad (61)\\
&= \sum_{t=1}^{T_{\text{stop}}} \left\langle \boldsymbol{\theta}_t \ , \ \underbrace{\mathbb{E}_{X_t}[W_*^\intercal \mathbf{x}_t(a_t)|H_{t-1}]}_{\mathbf{v}_t} - \underbrace{\mathbb{E}_{X_t}[W_*^\intercal X_t\boldsymbol{\pi}^*(X_t)]}_{\leqslant \frac{B\mathbf{1}}{T} \text{ by definition of } \boldsymbol{\pi}^*} \right\rangle \\
&\geqslant \sum_{t=1}^{T_{\text{stop}}} \left\langle \boldsymbol{\theta}_t \ , \ \mathbf{v}_t - \mathbf{1}\frac{B}{T} \right\rangle = \sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t) \ , \qquad (62)
\end{aligned}$$

where $g_t(\boldsymbol{\theta}_t)$ is the objective maximized by OCO algorithm in each step. We therefore obtain lower bounds on $\sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t)$. Let $\boldsymbol{\theta}^* = \underset{\|\boldsymbol{\theta}\|_1 \leqslant 1, \boldsymbol{\theta}_i \geqslant 0}{\text{argmax}} \sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta})$. If $R_D(T_{\text{stop}})$ is the regret of the OCO algorithm up to time T_{stop}, the following is true,

$$\sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t) \geqslant \sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}^*) - R_D(T_{\text{stop}}) \ .$$

Consider the two cases, when $T_{\text{stop}} < T$ and $T_{\text{stop}} = T$.

Case 1: If $T_{\text{stop}} < T$, then $\sum_{t=1}^{T_{\text{stop}}} \langle \mathbf{e}_j, \mathbf{v}_t\rangle \geqslant B' = B - \beta T_0$ for some $j = 1, \ldots, d$. Therefore,

$$\begin{aligned}
\sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t) \geqslant \sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}^*) - R_D(T_{\text{stop}}) \geqslant \sum_{t=1}^{T_{\text{stop}}} g_t(\mathbf{e}_j) - R_D(T_{\text{stop}}) = \sum_{t=1}^{T_{\text{stop}}} \left\langle \mathbf{e}_j, \mathbf{v}_t - \mathbf{1}\frac{B}{T} \right\rangle - R_D(T_{\text{stop}}) \geqslant \\
B\left(1 - \frac{T_{\text{stop}}}{T}\right) - \beta T_0 - R_D(T_{\text{stop}}) \ .
\end{aligned}$$

Case 2: If $T_{\text{stop}} = T$, then $B\left(1 - \frac{T_{\text{stop}}}{T}\right) = 0$. Therefore,

$$\sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t) \geqslant \sum_{t=1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}^*) - R_D(T_{\text{stop}}) \geqslant \sum_{t=1}^{T_{\text{stop}}} g_t(\mathbf{0}) - R_D(T_{\text{stop}}) = 0 - R_D(T_{\text{stop}}) = B\left(1 - \frac{T_{\text{stop}}}{T}\right) - R_D(T_{\text{stop}}) \ .$$

Therefore, from the results for case 1 and case 2, we get,

$$\sum_{t=1}^{T_{\text{stop}}} g_t(\theta_t) \geqslant B\left(1 - \frac{T_{\text{stop}}}{T}\right) - \beta T_0 - R_D(T_{\text{stop}}) \ . \qquad (63)$$

Therefore from (62) and (63) we get the following advertised result,

$$\mathbb{E}_X[\gamma_a(\tau) - \gamma_\pi(\tau)] \geqslant B\left(1 - \frac{T_{\text{stop}}}{T}\right) - \beta T_0 - R_D(T_{\text{stop}}) \ .$$

■

From Proposition 1 we get with probability atleast $1 - \delta$,

$$R_D(T_{\text{stop}}) = O(\sqrt{T(\log d)}) \leqslant O(R(T)) \ . \qquad (64)$$

Also from result of Lemma 1 with probability atleast $1 - \delta$, β is bounded with the assumption $\sigma = \Omega(1/\sqrt{m})$.

$$\beta \leqslant O\left(1 + \sigma\sqrt{m\log(TK/\delta)}\right) \qquad (65)$$

From result of Lemma 4, we have with probability atleast $1-\delta$,

$$\frac{\mathrm{OPT}_1}{B} \leqslant Z(T_0) \leqslant \left(\frac{\mathrm{OPT}_1}{B}+1\right) . \tag{66}$$

Therefore, from equations (58), (59), (64), (65), (66) and the result of Lemma 10, with probability atleast $1-5\delta$

$$\mathbb{E}_X[\mathrm{REW}] \geqslant \frac{T_{\text{stop}}}{T}\,\mathrm{OPT}_1 - ZB\left(\frac{T_{\text{stop}}}{T}-1\right) - O\left(\left(\frac{\mathrm{OPT}_1}{B}+1\right)\beta T_0\right) - O\left(\left(\frac{\mathrm{OPT}_1}{B}+1\right)R(T)\right) .$$

$$\mathbb{E}_X[\mathrm{REW}] \geqslant \mathrm{OPT}_1 - O\left(\left(\frac{\mathrm{OPT}_1}{B}+1\right)\beta T_0\right) - O\left(\left(\frac{\mathrm{OPT}_1}{B}+1\right)R(T)\right) , \tag{67}$$

which is the result in expectation. To bound the actual total reward we use the result of Corollary 3 with $z_t = \langle \mathbf{x}_t(a_t) - \mathbb{E}_X[\mathbf{x}_t(a_t)], \boldsymbol{\mu}_* \rangle$. The vector $\mathbf{x}_t(a_t) - \mathbb{E}_X[\mathbf{x}_t(a_t)]$ is a sub-Gaussian random vector with sub-Gaussian norm $\|\mathbf{x}_t(a_t) - \mathbb{E}_X[\mathbf{x}_t(a_t)]\|_{\psi_2} = \sup_{u \in S^{m-1}} \|\langle \mathbf{x}_t(a_t) - \mathbb{E}_X[\mathbf{x}_t(a_t)], u\rangle\|_{\psi_2} \leqslant c_1(1+\sigma)$. Therefore $z_t = \langle \mathbf{x}_t(a_t) - \mathbb{E}_X[\mathbf{x}_t(a_t)], \boldsymbol{\mu}_*\rangle$ is a $c_2(1+\sigma)$ sub-Gaussian random variable. Therefore applying Corollary 3 we get with probability atleast $1-\delta$,

$$\mathrm{REW} \geqslant \mathbb{E}_X[\mathrm{REW}] - O(\sqrt{T\log(1/\delta)}) \geqslant \mathbb{E}_X[\mathrm{REW}] - O(R(T)) . \tag{68}$$

Combining equations (67), (68) gives the advertised result. ■

F. Regret for Smoothed Adversarial Linear Bandits with Knapsacks

Theorem 3 *Assume $B \geqslant \Omega(T^{3/4})$. Then with probability at least $1-6\delta$ following is a lower bound on the reward obtained by the greedy algorithm*

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}_2}{16d\lceil\log T\rceil} - O\left(\left(\frac{\mathrm{OPT}_2}{B}+1\right)R(T)\right) , \tag{25}$$

where $R(T) = O\left(\frac{\left(w(A)+\sqrt{\log(Td)/\delta}\right)\beta\sqrt{T}\log K}{\sigma}\right)$.

Proof: We first show that by design resource consumption is not exceeded over the T time steps. The algorithm allocates a budget $B_0 = B'/2\lceil\log T\rceil$ to each epoch. $\lceil\log T\rceil$ is the maximum number of epochs, so a budget of $B'/2$ is allocated for running the greedy algorithm. The other $B'/2$ budget is allocated for the random exploration rounds in each epoch once any of the resource consumption exceeds B_0. In the random exploration rounds an arm is chosen uniformaly at random with probability $T^{-1/4}$. Therefore with T rounds with probability $1-\delta$ at most $T^{3/4} + 3\sqrt{T^{3/4}\log(1/\delta)}$ rounds are random exploration rounds. Therefore if $B > 4T^{3/4}$ resource consumptions are not exceeded over the T time steps.

We now focus on the last completed epoch of the algorithm from time step T_1 to T_2. Let κ be the value of parameter j in the algorithm during time step T_1. Therefore, we have the following relationships with $\widehat{\mathrm{OPT}}(T_1;\omega(T_1)), \widehat{\mathrm{OPT}}(T_2;\omega(T_2)), \widehat{\mathrm{OPT}}(T;\omega(T))$ denoting the maximum reward that can be obtained using only contexts before time steps T_1, T_2, T respectively.

$$\begin{aligned}
&\widehat{\mathrm{OPT}}(T;\omega(T)) \leqslant 2^{\kappa+2} \quad \text{since a new epoch was not started after } T_2 \\
&\widehat{\mathrm{OPT}}(T_2;\omega(T_2)) \geqslant 2^{\kappa+1} \quad \text{a new epoch begins at time step } T_2 \\
&2^{\kappa} \leqslant \widehat{\mathrm{OPT}}(T_1;\omega(T_1)) < 2^{\kappa}+2\beta \quad \text{a new epoch begins at time step } T_1\text{, reward bounded by } \beta .
\end{aligned} \tag{69}$$

Also using estimation error bounds for $\widehat{\mathrm{OPT}}(T;\omega(T))$, $\widehat{\mathrm{OPT}}(T_1;\omega(T_1))$ and $\widehat{\mathrm{OPT}}(T_2;\omega(T_2))$ from Lemma 4 with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{OPT}(T;\omega(T)) - 2R(T) \leqslant \widehat{\mathrm{OPT}}(T;\omega(T)) &\leqslant \mathrm{OPT}(T;\omega(T)) + 9R(T)\left(\frac{\mathrm{OPT}(T;\omega(T))}{B}+1\right) \\
\mathrm{OPT}(T_1;\omega(T_1)) - 2R(T) \leqslant \widehat{\mathrm{OPT}}(T_1;\omega(T_1)) &\leqslant \mathrm{OPT}(T_1;\omega(T_1)) + 9R(T)\left(\frac{\mathrm{OPT}(T_1;\omega(T_1))}{B}+1\right) \\
\mathrm{OPT}(T_2;\omega(T_2)) - 2R(T) \leqslant \widehat{\mathrm{OPT}}(T_2;\omega(T_2)) &\leqslant \mathrm{OPT}(T_2;\omega(T_2)) + 9R(T)\left(\frac{\mathrm{OPT}(T_2;\omega(T_2))}{B}+1\right) ,
\end{aligned} \tag{70}$$

where $\mathrm{OPT}(T;\omega(T)), \mathrm{OPT}(T_1;\omega(T_1)), \mathrm{OPT}(T_2;\omega(T_2))$ are the optimal values computed at time step T, T_1, T_2 using observed data before T, T_1, T_2 respectively . Also $Z(T_1)$ for the last completed epoch is set at the beginning of round T_1.

$$Z(T_1) = \frac{\widehat{\mathrm{OPT}}(T_1;\omega(T_1))}{2dB} \; . \tag{71}$$

We now focus on analyzing the performance of the algorithm during the last completed epoch. Let $T_{\text{stop}} \leqslant T_2$ denote the stopping time. $T_{\text{stop}} = T_2$ if no resource is exhausted else the first time step when any resource consumption exceeds the allocation B_0.

Case 1: $T_{\textbf{stop}} < T_2$

By Theorem 1 and 9,

$$\sum_{t=T_1}^{T_{\text{stop}}} \langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) \rangle \geqslant \gamma_a(T_1, T_{\text{stop}}) - O(Z(T_1)R(T)) \; , \tag{72}$$

where $\gamma_a(T_1, T_{\text{stop}}) = \sum_{t=T_1}^{T_{\text{stop}}} \langle W_*^\intercal \mathbf{x}_t(a_t), \boldsymbol{\theta}_t \rangle$. Since $\mathbb{E}[\mathbf{v}_t(a_t) | X_t, a_t, H_{t-1}] = W_*^\intercal \mathbf{x}_t(a_t)$ and therefore by Azuma-Hoeffding $\| \sum_{t=T_1}^{T_{\text{stop}}} \mathbf{v}_t(a_t) - W_*^\intercal \mathbf{x}_t(a_t) \|_\infty \leqslant R(T)$. Therefore to bound $\gamma_a(T_{\text{stop}})$ we will bound $\langle \mathbf{v}_t(a_t), \boldsymbol{\theta}_t \rangle$. For the dual OCO algorithm with $g_t(\boldsymbol{\theta}_t) = \langle \boldsymbol{\theta}_t, (\mathbf{v}_t(a_t) - \frac{B_0}{T}\mathbf{1}) \rangle$, let $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_1 = 1, \boldsymbol{\theta}_i \geqslant 0}{\operatorname{argmax}} \sum_{t=T_1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta})$. Therefore, due to the OCO algorithm bounds with probability at least $1 - \delta$,

$$\sum_{t=T_1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}_t) \geqslant \sum_{t=T_1}^{T_{\text{stop}}} g_t(\boldsymbol{\theta}^*) - R_D(T_{\text{stop}}) \;\; \Rightarrow \;\; \sum_{t=T_1}^{T_{\text{stop}}} \langle \mathbf{v}_t(a_t), \boldsymbol{\theta}_t \rangle \geqslant \sum_{t=T_1}^{T_{\text{stop}}} \langle \mathbf{v}_t(a_t), \boldsymbol{\theta}^* \rangle - R_D(T_{\text{stop}}) \; , \tag{73}$$

where $R_D(T_{\text{stop}})$ is the dual OCO algorithm regret. Now since θ^* maximizes the objective, for some $1 \leqslant i \leqslant d$,

$$\sum_{t=T_1}^{T_{\text{stop}}} \langle \mathbf{v}_t(a_t), \boldsymbol{\theta}^* \rangle \geqslant \sum_{t=T_1}^{T_{\text{stop}}} \langle \mathbf{v}_t(a_t), \mathbf{e}_i \rangle \geqslant B_0 \; . \tag{74}$$

Since $R_D(\tau) \leqslant R(T)$ from equation (72) we get the following,

$$\sum_{t=T_1}^{T_{\text{stop}}} \langle \boldsymbol{\mu}_*, \mathbf{x}_t(a_t) \rangle \geqslant Z(T_1)B_0 - O(Z(T_1)R(T)) \; . \tag{75}$$

Case 2: $T_{\textbf{stop}} = T_2$

Denote the total optimal reward by $\mathrm{OPT}(T_1, T_2; \omega(T_2))$ which is the optimal reward between time steps T_1 and T_2 computed with reference to time step T_2 using all observed data before T_2. By Theorem 1 and 9,

$$\mathrm{REW} = \mathrm{OPT}(T_1, T_2; \omega(T_2)) - Z(T_1)\gamma_1(T_2) + Z(T_1)\gamma_2(T_2) - O(Z(T_1)R(T)) \; . \tag{76}$$

The following is always true,

$$\gamma_1(T_2) = \sum_{t=T_1}^{T_2} \langle W_* X_t \pi^*(X_t; \omega(T_2)), \boldsymbol{\theta}_t \rangle \leqslant dB \; . \tag{77}$$

Also applying OCO regret bounds with probability atleast $1 - \delta$,

$$\gamma_1(T_2) \geqslant -R_D(T_2) \; , \tag{78}$$

with $R_D(T_2) = O(R(T))$. Therefore from (76), (77) and (78), we have

$$\mathrm{REW} \geqslant \mathrm{OPT}(T_1, T_2; \omega(T_2)) - Z(T_1)dB - O(Z(T_1)R(T)) \; . \tag{79}$$

Combining results (75) and (79),

$$\mathrm{REW} \geqslant \min(\mathrm{OPT}(T_1, T_2; \omega(T_2)) - Z(T_1)dB, Z(T_1)B_0) - O(Z(T_1)R(T))\;. \tag{80}$$

Next we focus on expressing $Z(T_1), \mathrm{OPT}(T_1, T_2; \omega(T_2))$ in terms of the optimal reward $\mathrm{OPT} = \mathrm{OPT}(T; \omega(T))$ and budget B. Now consider the quantity $\mathrm{OPT}(T_1, T_2; \omega(T_2)) - Z(T_1)dB$.

$$\begin{aligned}
\mathrm{OPT}(T_1, T_2; \omega(T_2)) - Z(T_1)dB &\geqslant \mathrm{OPT}(T_2; \omega(T_2)) - \mathrm{OPT}(T_1; \omega(T_2)) - Z(T_1)dB \\
&\geqslant \mathrm{OPT}(T_2; \omega(T_2)) - \widehat{\mathrm{OPT}}(T_1; \omega(T_1)) - O(R(T)) - \frac{\widehat{\mathrm{OPT}}(T_1; \omega(T_1))}{2} \\
&\qquad\qquad\qquad\qquad \text{....use (71), upper bound for } \widehat{\mathrm{OPT}}(T_1; \omega(T_1)) \text{ from (70)} \\
&\geqslant \widehat{\mathrm{OPT}}(T_2; \omega(T_2)) - O\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) R(T)\right) - \frac{3}{2}\widehat{\mathrm{OPT}}(T_1; \omega(T_1)) - O(R(T)) \\
&\qquad\qquad\qquad\qquad \text{....use lower bound for } \mathrm{OPT}(T_2) \text{ from (70)} \\
&\geqslant 2^{\kappa+1} - \frac{3}{2}2^{\kappa} - O\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) R(T)\right) \\
&\qquad\qquad\qquad\qquad \text{....substituting bounds from (69)} \\
&\geqslant 2^{\kappa+2}\left(\frac{1}{2} - \frac{3}{8}\right) - O\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) R(T)\right) \\
&\geqslant \frac{2^{\kappa+2} + O(R(T))}{8} - \frac{O(R(T))}{8} - O\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) R(T)\right) \\
&\geqslant \frac{\mathrm{OPT}}{8} - O\left(\left(\frac{\mathrm{OPT}}{B} + 1\right) R(T)\right) \\
&\qquad\qquad \text{....use upper bound for } \widehat{\mathrm{OPT}}(T; \omega(T)) \text{ from (69), then use (70)}
\end{aligned} \tag{81}$$

In the first line above the quantities $\mathrm{OPT}(T_1; \omega(T_2))$ and $\mathrm{OPT}(T_2; \omega(T_2))$ are the optimal reward at time steps T_1, T_2 with time step T_2 as reference..

$$\begin{aligned}
\pi^* &:= \operatorname*{argmax}_{\pi} \sum_{t=1}^{T_2} \boldsymbol{\mu}_*^{\intercal} X_t \pi(X_t; \omega(T_2)) \quad \text{s.t.} \quad \sum_{t=1}^{T_2} W_*^{\intercal} X_t \pi(X_t; \omega(T_2)) \leqslant B\mathbf{1} \\
\mathrm{OPT}(T_1; \omega(T_2)) &= \sum_{t=1}^{T_1} \boldsymbol{\mu}_*^{\intercal} X_t \pi^*(X_t; \omega(T_2)) \\
\mathrm{OPT}(T_2; \omega(T_2)) &= \sum_{t=1}^{T_2} \boldsymbol{\mu}_*^{\intercal} X_t \pi^*(X_t; \omega(T_2))
\end{aligned}$$

The first line above is true because we can always assume our algorithm ends at time T_2 choosing no-op arms after T_2. So $\mathrm{OPT}(T_2; \omega(T_2))$ is the maximum reward that can be obtained and reference optimal awards in all time steps before T_2 to time step T_2. Also $\widehat{\mathrm{OPT}}(T_1; \omega(T_1)) \leqslant 2^{\kappa} + 2\beta$ estimated at the end of time step T_1 is the maximum reward that can be obtained by using all resources and only using contexts before time step T_1 and hence $\widehat{\mathrm{OPT}}(T_1; \omega(T_2))$ computed at reference time T_2 is always less than or equal to $2^{\kappa} + 2\beta$.

Next consider quantity $Z(T_1)B_0$.

$$\begin{aligned}
Z(T_1)B_0 &= \frac{\widehat{\mathrm{OPT}}(T_1;\omega(T_1))}{2dB} \\
&\geqslant \frac{2^{\kappa}}{4d\lceil\log T\rceil} \text{ ... use lower bound for } \widehat{\mathrm{OPT}}(T_1;\omega(T_1)) \text{ from (69), } B_0 = B/2\lceil\log T\rceil \\
&\geqslant \frac{2^{\kappa+2}+O(R(T))}{16d\lceil\log T\rceil} - \frac{O(R(T))}{16d\lceil\log T\rceil} \\
&\geqslant \frac{2^{\kappa+2}+O(R(T))}{16d\lceil\log T\rceil} - \frac{O(R(T)}{16d\lceil\log T\rceil} \\
&\geqslant \frac{\mathrm{OPT}}{16d\lceil\log T\rceil} - O(R(T))\ . \qquad (82)
\end{aligned}$$

Also,

$$Z(T_1) = \frac{\widehat{\mathrm{OPT}}(T_1;\omega(T_1))}{2dB} \leqslant O\left(\frac{\mathrm{OPT}}{B}+1\right)\ . \qquad (83)$$

Therefore from equations (80), (81), (82), (83), we get

$$\mathrm{REW} \geqslant \frac{\mathrm{OPT}}{16d\lceil\log T\rceil} - O\left(\left(\frac{\mathrm{OPT}}{B}+1\right)R(T)\right)\ . \qquad (84)$$