Linear Contextual Bandits with Global Constraints and Objective

Shipra Agrawal*       Nikhil R. Devanur†

July 27, 2015

Abstract

We consider the linear contextual bandit problem with global convex constraints and a concave objective function. In each round, the outcome of pulling an arm is a vector, that depends linearly on the context of that arm. The global constraints require the average of these vectors to lie in a certain convex set. The objective is a concave function of this average vector. This problem turns out to be a common generalization of classic linear contextual bandits (linContextual) [8, 17, 1], bandits with concave rewards and convex knapsacks (BwCR) [4], and the online stochastic convex programming (OSCP) problem [5]. We present algorithms with near-optimal regret bounds for this problem. Our bounds compare favorably to results on the unstructured version of the problem [6, 12] where the relation between the contexts and the outcomes could be arbitrary, but the algorithm only competes against a fixed set of policies. We combine techniques from the work on linContextual, BwCR and OSCP in a nontrivial manner while also tackling new difficulties that are not present in any of these special cases.

*Microsoft Research. ashipra@gmail.com.
†Microsoft Research. nikdev@microsoft.com.
1 Introduction

In the contextual bandit problem [8, 14, 22, 3], the decision maker observes a sequence of contexts (or features). In every round she needs to pull one out of $K$ arms, after observing the context for that round. The outcome of pulling an arm may be used along with the contexts to decide future arms. Contextual bandit problems have found many useful applications such as online recommendation systems, online advertising, and clinical trials, where the decision in every round needs to be customized to the features of the user being served. The linear contextual bandit problem [1, 8, 17] is a special case of the contextual bandit problem, where the outcome is linear in the feature vector encoding the context. As pointed by [3], contextual bandit problems represent a natural half-way point between supervised learning and reinforcement learning: the use of features to encode contexts and the models for the relation between these feature vectors and the outcome are often inherited from supervised learning, while managing the exploration-exploitation tradeoff is necessary to ensure good performance in reinforcement learning. The linear contextual bandit problem can thus be thought of as a midway between the linear regression model of supervised learning, and reinforcement learning.

Recently, there has been a significant interest in introducing multiple “global constraints” in the standard bandit setting [11, 4, 12, 6]. Such constraints are crucial for many important real-world applications. For example, in clinical trials, the treatment plans may be constrained by the total availability of medical facilities, drugs and other resources. In online advertising, there are budget constraints that restrict the number of times an ad is shown. Other applications include dynamic pricing, dynamic procurement, crowdsourcing, etc.; see [11, 4] for many such examples.

In this paper, we consider a very general version of the linear contextual bandit problem with global constraints and objective (henceforth, linCBwCR). In this problem, the context vectors are generated i.i.d. in every round from some unknown distribution, and the expected outcome on picking an arm is in the form of a vector that depends linearly on the context vector. The aim of the decision maker is to maximize a given concave objective function of the average of these outcome vectors while ensuring that the average lies in a given convex set. Below, we give a more precise definition of this problem. We use the following notational convention throughout: vectors are denoted by bold face lower case letters, while matrices are denoted by regular face upper case letters. Other quantities such as sets, scalars, etc. may be of either case, but never bold faced. All vectors are column vectors, i.e., a vector in $n$ dimensions is treated as an $n \times 1$ matrix. The transpose of matrix $A$ is $A^\top$.

**Definition 1 (linCBwCR).** There are $K$ “arms”, which we identify with the set $[K]$. The algorithm is initially given as input a convex set $S \subseteq [0,1]^d$, and an $L$-Lipschitz concave function $f$ with domain $[0,1]^d$. We then proceed in rounds $t = 1, 2, \ldots, T$, where in every round $t$, a pair of a context matrix and an outcome matrix $(X_t, V_t) \in [0,1]^{m \times K} \times [0,1]^{d \times K}$ are drawn from an unknown distribution $\mathcal{D}$, independent of everything in previous rounds. The column of $X_t$ corresponding to arm $a \in [K]$ is called the context vector for arm $a$ and is denoted by $x_t(a) \in [0,1]^m$. The column of $V_t$ corresponding to an arm $a$ is called the “outcome” vector for arm $a$ and is denoted by $v_t(a) \in [0,1]^d$. There exist an (unknown) weight matrix $W_\ast \in [0,1]^{m \times d}$ such that

$$
\mathbb{E}[V_t|X_t] = W_\ast^\top X_t.
$$

In each round, the algorithm first observes context $X_t$, and then chooses an arm $a_t \in A$, and finally observes $v_t(a_t)$. The goal of the algorithm is to pick arms such that the average vector $\frac{1}{T} \sum_{t=1}^T v_t(a_t)$ maximizes $f(\frac{1}{T} \sum_{t=1}^T v_t(a_t))$ and is inside $S$.

We compare the performance of an algorithm to that of the optimal adaptive policy that knows the distribution $\mathcal{D}$ and the weight matrix $W_\ast$, and can take into account the history upto that point as well as the current context to decide (possibly with randomization) which arm to pull at time $t$. However, it is easier to work with an upper bound on this, which is the optimal expected reward of a static policy
that is required to satisfy the constraints only in expectation. This technique has been used in several related problems and is standard by now [20, 11].

**Definition 2 (Optimal Static Policy).** We consider policies that are context dependent but are non-adaptive: for a policy \( q \) let \( q(X) \in \Delta^K \) (the unit simplex) denote the probability distribution over arms played when the context is \( X \in \mathcal{X} \). Define \( \mathbf{v}(q) \) to be the expected outcome vector of policy \( q \), i.e.

\[
\mathbf{v}(q) := \mathbb{E}_{(X,V) \sim \mathcal{D}}[Vq(X)] = \mathbb{E}_X[W^T Xq(X)].
\]

Let \( q^* := \arg \max_q f(\mathbf{v}(q)) \) such that \( \mathbf{v}(q) \in S \) be the optimal static policy. We assume that there exists a feasible policy, i.e. \( \mathbf{v}(q^*) \in S \), and we denote the optimal value by \( \text{OPT} := f(\mathbf{v}(q^*)) \).

**Definition 3 (Regret).** We define two regrets: regret in objective and regret in constraints. Let \( a_t \) be the arm played at time \( t \) by the algorithm. For convenience of notation let us define \( \bar{\mathbf{v}}_{1:T} := \frac{1}{T} \sum_{t=1}^T \mathbf{v}_t(a_t) \),

- Regret in objective. \( \text{avg-regret}^1(T) = \text{OPT} - f(\bar{\mathbf{v}}_{1:T}) \).

- Regret in constraints is defined as the distance of the average of played vectors from \( S \); we let \( d(\cdot, \cdot) \) denote a distance function. \( \text{avg-regret}^2(T) := d(\bar{\mathbf{v}}_{1:T}, S) \).

We consider following useful special cases.

**Feasibility problem (linCBwC):** In this special case of linCBwCR, there is no objective function \( f \), and the aim of the algorithm is to have the average outcome vector \( \bar{\mathbf{v}}_{1:T} \) be in the set \( S \). The performance of the algorithm is measured by the distance of \( \bar{\mathbf{v}}_{1:T} \) to \( S \), i.e., by \( \text{avg-regret}^2(T) \). We will first illustrate our algorithm and proof techniques for this special yet nontrivial case.

**Budget constraints (linCBwK):** This is a well studied special case with linear objectives and constraints; a generalization of the Bandits with Knapsacks (BwK) problem of [11]. Here, the outcome vector can be broken down into two components: a scalar reward \( r_t(a_t) \in [0, 1] \) and a \( d-1 \)-dimensional consumption vector \( \mathbf{v}_t(a_t) \in [0, 1]^{d-1} \). The objective is to maximize \( \sum_{t=1}^T r_t(a_t) \) while ensuring that \( \sum_{t=1}^T \mathbf{v}_t(a_t) \leq B \mathbf{1} \), where \( \mathbf{1} \) is the vector of all 1s and \( B > 0 \) is some scalar. Such budget constraints are equivalent to having the constraint set \( S \) be equal to \( \{ \mathbf{v} : 0 \leq \mathbf{v} \leq B \mathbf{1} \} \). Also, it is possible to simply stop once the budget is consumed; equivalently we assume that the algorithm always has the “do nothing” option of getting 0 reward and 0 cost. For this problem, we provide an algorithm that always satisfies the budget constraints, i.e., \( \text{avg-regret}^2(T) = 0 \), while bounding \( \text{avg-regret}^1(T) \).

### 1.1 Main results

Our main result is an algorithm with near-optimal regret bound for linCBwCR. The algorithm needs to know a certain problem dependent parameter \( Z \), defined by Assumption 1 in Section 4. Such knowledge is not required for linCBwC, however. For the special case of budget constraints (linCBwK), any \( Z \geq 2 \text{OPT} \frac{m}{T} \) can be used, and we show how one such \( Z \) can be obtained from the first few rounds.

Below we state the main result when \( \| \cdot \| \) denotes either \( \ell_2 \) or \( \ell_\infty \) norm. In the text, we will provide a more detailed theorem statement that is applicable to any norm.

**Theorem 1.** Let \( \| \cdot \| \) denote either \( \ell_2 \) or \( \ell_\infty \) norm, and be used to define the distance \( d(\cdot, \cdot) \). Assume \( f \) is \( L \)-Lipschitz with respect to \( \| \cdot \| \). Then, for linCBwCR, there is an algorithm that takes as input a problem-dependent parameter \( Z \geq L \) (satisfying Assumption 1), and with probability at least \( 1 - \delta \) has

\[
\text{avg-regret}^1(T) = O \left( Zm\|1_d\| \sqrt{\ln(dT/\delta) \ln(T)/T} \right), \quad \text{avg-regret}^2(T) = O \left( m\|1_d\| \sqrt{\ln(dT/\delta) \ln(T)/T} \right).
\]
For linCBwC, no knowledge of $Z$ is needed, and the algorithm has the same avg-regret$^2(T)$ as above.

Remark 1. (Near-optimality of regret bounds) In [18], it was shown that for the linear contextual bandits problem, no online algorithm can achieve a regret bound better than $\Omega(m\sqrt{T})$. In fact, they prove this lower bound for linear contextual bandits with static contexts. Since that problem is a special case of our linCBwKR problem with $d = 1$, this shows that the dependence on $m$ and $T$ in the above (average) regret bounds is optimal up to log factors. The dependence on $Z||1_d||$ is common in special cases of this problem [4, 5]. Moreover, we don’t need to know $D$, as is assumed in related work [6, 12, 38].

Remark 2. (Budget constraints) For the special case of linCBwK, we show how to estimate $Z$ using the observations from the first few rounds. We get an overall regret bound of $\tilde{O}((OPT/(B/T) + 1)^{m/\sqrt{T}})$, which compares favorably to the regret bounds for the general contextual version obtained in [12, 6]. (Details in Section 1.2.) The only difference is that we need the budget $B$ to be $\Omega(mT^{3/4})$. Getting similar bounds for budgets as small as $B = \Theta(m\sqrt{T})$ is an interesting open problem.

Due to space constraints, we have eliminated many proofs from the main text. All the missing proofs are in the appendix.

1.2 Challenges and related work

Relation to BwCR [4] and OSCP [5]. It is easy to see that the linCBwKR problem is a generalization of the linear contextual bandits problem. There, the outcome is scalar and the goal is to simply maximize the sum of these. Remarkably, the linCBwKR problem also turns out to be a common generalization of bandits with concave rewards and convex knapsacks (BwCR) problem considered in [4], and the online stochastic convex programming (OSCP) problem of [5]. In both BwCR and OSCP problems, the outcome of every round $t$ is a vector $v_t$ and the goal of the algorithm is to maximize $f(\frac{1}{T} \sum_{t=1}^T v_t)$ while ensuring that $\frac{1}{T} \sum_{t=1}^T v_t \in S$. The problems differ in how this vector is picked. In the OSCP problem, in every round $t$, the algorithm may pick any vector from a given set $A_t$ of $d$-dimensional vectors. The set $A_t$ is drawn i.i.d. from an unknown distribution over sets of vectors. This corresponds to the special case of linCBwKR, where $m = d$ and the outcome $V_t$ is always equal to the context $X_t$. In the BwCR problem, there is a fixed set of arms, and for each arm there is an unknown distribution over vectors. The algorithm picks an arm and a vector is drawn from the corresponding distribution for that arm. This corresponds to the special case of linCBwKR, where $m = K$ and the context $X_t = I$, the identity matrix, for all $t$.

We use techniques from all three special cases: our algorithms follow the primal-dual paradigm using Fenchel duality, that was established in [4]. In order to deal with linear contexts, we use techniques from [1, 8, 17] to estimate the weight matrix $W_*$, and define “optimistic estimates” of $W_*$. We also use the technique of combining the objective and the constraints using a certain tradeoff parameter that was introduced in [5]. Further new difficulties arise, such as in estimating the optimum value from the first few rounds, a task that follows from standard techniques in each of the special cases but is very challenging here. One can see that the problem is indeed more than the sum of its parts, from the fact that we get the optimal bound for linCBwK only when $B \geq \tilde{O}(mT^{3/4})$, unlike either special case for which the optimal bound holds for all $B$ (but is meaningful only for $B = \tilde{O}(m\sqrt{T})$).

The approach in [4] (for BwCR) extends to the case of “static” contexts, where each arm has a context that doesn’t change over time. The OSCP problem of [5] is not a special case of linCBwKR with static contexts; this is one indication of the additional difficulty of dynamic over static contexts.

Relation to general contextual bandits. There have been recent papers [6, 12] that solve problems similar to linCBwKR but for general contextual bandits. Here the relation between contexts and

---

1It was incorrectly claimed in [4] that the approach can be extended to dynamic contexts without much modifications.
outcome vectors is arbitrary and the algorithms compete with an arbitrary fixed set of context dependent policies $\Pi$, with the regret bounds scaling as $\sqrt{K \log(|\Pi|)}$. (In our results, this $\sqrt{K \log(|\Pi|)}$ is replaced by $m$.) Since linCBwCR is a special case of this problem, we may compare the regret bounds obtained by applying these results. Firstly, we have no dependence on $K$, which enables us to consider problems with large action spaces. Further, with global constraints and objective, the effective policy space for linear bandits is \textit{doubly exponential} in its dimensions, thus giving exponentially large regret bounds for linCBwCR using this approach. The key is to determine the size of the smallest policy set that is guaranteed to contain a near optimal policy. Trivially, the number of possible context dependent policies $K_{|X|}$, where $X$ denotes the space of contexts. In linCBwCR, $X = [0, 1]^{m \times K}$, giving $|X| = O(1/\epsilon^{(mK)})$ using a discretization of the space into multiples of some small enough $\epsilon$.

We could try to reduce the policy space using linearity. In linear contextual bandits, the objective which is the sum of rewards, is decomposable across time steps. Given the weight vector $W_*$, the optimal policy is to chose the arm $a$ that maximizes the expected reward $W_*^\top x_t(a)$ at time $t$, which is independent of the distribution $D$ of the contexts. It thus suffices to have $\Pi$ be the set consisting of one policy of this type for every possible $W_*$, giving $|\Pi| = \frac{1}{m} \left( \begin{array}{c} mK \\ d \end{array}\right)$ (since $d = 1$). However, in the linCBwCR problem, due to nondecomposable objective $f(\frac{1}{T} \sum_{t} W_*^\top X_t q(X_t))$, and/or distance function $d(\frac{1}{T} \sum_{t} W_*^\top X_t q(X_t), S)$, the optimal policy cannot locally maximize the reward at every step. Even for a fixed $W_*$, the optimal policy may depend on $D$. We can have $\Pi$ consist of one policy for every possible pair of $D$ and $W_*$, but that gives $|\Pi| = |X| \cdot \frac{1}{m}$ which is again \textit{doubly exponential} in $m \times K$ since $|X| = O(1/\epsilon^{(mK)})$.

\textbf{Other related work.} Budget constraints in a bandit setting has received considerable attention, but most of the early work focussed on special cases such as a single budget constraint in the regular (non-contextual) setting [21, 24, 26, 29, 35, 36]. Recently, [38] showed an $O(\sqrt{T})$ regret in the linear contextual setting with a single budget constraint, when costs depend only on contexts and not arms. Budget constraints that arise in particular applications such as online advertising [15, 31], dynamic pricing [9, 13] and crowdsourcing [10, 33, 34] have also been considered. There has also been a long line of work studying special cases of the OSCP problem [19, 20, 23, 7, 28, 25, 37, 30, 27, 16].

\section{Preliminaries}

\subsection{Confidence Ellipsoid}

Consider a stochastic process which in each round $t$, generates a pair of observations $(r_t, y_t)$, such that $r_t$ is an unknown linear function of $y_t$ plus some 0-mean bounded noise, i.e., $r_t = \mu_*^\top y_t + \eta_t$, where $y_t, \mu_* \in \mathbb{R}^m$, $|\eta_t| \leq 2R$, and $E[\eta_t | y_1, r_1, \ldots, y_{t-1}, r_{t-1}, y_t] = 0$.

At any time $t$, a high confidence estimate of the unknown vector $\mu_*$ can be obtained by building a “Confidence Ellipsoid” around the $\ell_2$-regularized least-square estimate $\hat{\mu}_t$ constructed from the observations made so far. This technique is common in prior work on linear contextual bandits (e.g., in [8, 17, 1]). For any regularization parameter $\lambda > 0$, let

$$M_t := \lambda I + \sum_{i=1}^{t-1} y_i y_i^\top, \quad \hat{\mu}_t := M_t^{-1} \sum_{i=1}^{t-1} y_i r_i.$$ 

The following result from [1] shows that $\mu_*$ lies with high probability in an ellipsoid with center $\hat{\mu}_t$. For any positive semi-definite (PSD) matrix $M$, define the $M$-norm as $\|\mu\|_M := \sqrt{\mu^\top M \mu}$. The confidence ellipsoid at time $t$ is defined as

$$C_t := \left\{ \mu \in \mathbb{R}^m : \|\mu - \hat{\mu}_t\|_{M_t} \leq R \sqrt{m \ln \left( \frac{(1+tm)/\lambda}{\delta} \right) + \sqrt{\lambda m}} \right\}.$$ 

\textbf{Lemma 1 (Theorem 2 of [1]).} If $\forall t$, $\|\mu_*\|_2 \leq \sqrt{m}$ and $\|y_t\|_2 \leq \sqrt{m}$, then with prob. $1 - \delta$, $\mu_* \in C_t$.  

4
Another useful observation about this construction is stated below. It first appeared as Lemma 11 of [8], and was also proved as Lemma 3 in [17].

**Lemma 2** (Lemma 11 of [8]). \( \sum_{t=1}^{T} \| y_t \|_{M_t^{-1}} \leq \sqrt{mT \ln(T)}. \)

As a corollary of the above two lemmas, we obtain a bound on the total error in the estimate provided by “any point” from the confidence ellipsoid. (Proof is in Appendix B.)

**Corollary 1.** For \( t = 1, \ldots, T, \) let \( \hat{\mu}_t \in C_t \) be a point in the confidence ellipsoid, with \( \lambda = 1, 2R = 1. \) Then, with probability \( 1 - \delta, \)

\[
\sum_{t=1}^{T} |\hat{\mu}_t^T y_t - \mu^*_t y_t| \leq 2m \sqrt{T \ln((1+Tm)/\delta) \ln(T)}.
\]

### 2.2 Fenchel duality

As mentioned earlier, our algorithms are primal-dual algorithms, that use Fenchel duality. Below we provide the basic background on this useful concept. Let \( h \) be a convex function defined on \([0, 1]^d\). We define \( h^* \) as the Fenchel conjugate of \( h, \)

\[
h^*(\theta) := \max \{ y \cdot \theta - h(y) : y \in [0, 1]^d \}.
\]

Similarly for a concave function \( f \) on \([0, 1]^d\), define \( f^*(\theta) := \max_{y \in [0, 1]^d} \{ y \cdot \theta + f(y) \} \). Note that the Fenchel conjugates \( h^* \) and \( f^* \) are both convex functions of \( \theta \).

Suppose that at every point \( y \), every supergradient \( g_y \) of \( h \) (and \( f \)) have bounded dual norm \( \| g_y \|_* \leq L. \) Then, the following dual relationship is known between \( h \) and \( h^* \) (\( f \) and \( f^* \)).

**Lemma 3.** \( h(z) = \max_{\| \theta \|_* \leq L} \{ \theta \cdot z - h^*(\theta) \}, \ f(z) = \min_{\| \theta \|_* \leq L} \{ f^*(\theta) - \theta \cdot z \} \)

A special case is when \( h(y) = d(y, S) \) for some convex set \( S \). This function is 1-Lipschitz with respect to norm \( \| \cdot \| \) used in the definition of distance. In this case, \( h^*(\theta) = h_S(\theta) := \max_{y \in S} \theta \cdot y, \) and Lemma 3 specializes to the following relation (which also appears in [2]).

\[
d(y, S) = \max \{ \theta \cdot y - h_S(\theta) : \| \theta \|_* \leq 1 \}. \tag{3}
\]

### 2.3 Online Learning

The online convex optimization (OCO) problem considers a \( T \) round game played between a learner and an adversary, where in round \( t, \) the learner chooses a \( \theta_t \in \Omega, \) and then the adversary picks a concave function \( g_t(\theta_t) : \Omega \rightarrow \mathbb{R}. \) The learner’s choice \( \theta_t \) may only depend on learner’s and adversary’s choices in previous rounds, and we denote this by \( \theta_t := \text{OL-Predict}(\theta_1, g_1, \ldots, \theta_{t-1}, g_{t-1}). \) The goal of the learner is to minimize regret defined as the difference between the learner’s objective value and the value of the best single choice on hindsight:

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

Some popular algorithms for OCO are online mirror descent (OMD) and online gradient descent, which have very fast per step update rules, and provide the following regret guarantees.

**Lemma 4.** [32] The online mirror-descent algorithm for the OCO problem that achieves regret

\[
\mathcal{R}(T) = O(G \sqrt{DT}),
\]

where \( G \) is an upper bound on the norm of gradient of \( g_t \) at all time \( t, \) and \( D \) is the diameter of an appropriate regularizer (in the dual norm). The value of these parameters are problem specific. In particular, for \( g_t(\theta) \) of form \( g_t(\theta) = \theta \cdot z - h^*(\theta) \) and \( \Omega = \{ \theta : \| \theta \|_* \leq L \}, \) where \( h \) is an \( L \)-Lipschitz function, \( h^* \) is Fenchel dual of \( h, \) regret is bounded as \( \mathcal{R}(T) \leq O(L \sqrt{dT}) \) for \( \ell_2 \) norm, and \( O(L \sqrt{\log(d)T}) \) for \( \ell_\infty. \)
3 Algorithm for the Feasibility problem (linCBwC)

We start with the case when there are only constraints, i.e., we are only given the convex set $S$ and our goal is to minimize only avg-regret $^2(T)$. In every round $t$, our algorithm uses online learning to generate a vector $\theta_t$ that acts as the dual price to weigh different components of the outcome vector estimate. Using this weighted sum of outcome vector components, for every arm we generate an optimistic estimate of the weight matrix $W_*$, using techniques similar to the UCB based algorithms for linear contextual bandits. We then pick the best arm according to these optimistic estimates.

3.1 Optimistic estimates of weight matrix

Let $a_t$ denote the arm played by the algorithm at time $t$. In the beginning of every round, we use the outcome and contexts from previous rounds to construct a confidence ellipsoid for every column of $W_*$. For this, we use the techniques in Section 2.1 with $y_t = x_t(a_t)$, and $r_t = v_t(a_t)_j$ for every $j$.

As in Section 2.1, let $M_t := I + \sum_{i=1}^{t-1} x_t(a_i)x_t(a_i)^\top$, and construct the regularized least squares estimate for matrix $W_*$ as

$$\hat{W}_t := M_t^{-1} \sum_{i=1}^{t-1} x_t(a_i)v_t(a_i)^\top.$$ (4)

Let $w_j$ denote the $j^{th}$ column of matrix $W$. We define a confidence ellipsoid for each column $j$, as

$$C_{t,j} := \left\{ \mu \in \mathbb{R}^m : \|\mu - \hat{w}_{tj}\|_{M_t} \leq \sqrt{m \ln ((d+td)/\delta)} + \sqrt{m} \right\},$$

and denote by $G_t$, the Cartesian product of all these ellipsoids: $G_t := \{ W \in \mathbb{R}^{m \times d} : w_j \subseteq C_{t,j} \}$. Note that Lemma 1 implies $W_* \in G_t$ with probability $1 - \delta$. Now, given a vector $\theta_t \in \mathbb{R}^d$, we define the optimistic estimate of weight matrix at time $t$ w.r.t. $\theta_t$, for every arm $a \in [K]$, as:

$$\hat{W}_t(a) := \arg \min_{W \in G_t} x_t(a)^\top W \theta_t.$$ (5)

Using the definition of $\hat{W}_t$, along with the results in Lemma 1 and Corollary 1 about confidence ellipsoids, the following can be derived.

**Corollary 2.** With probability $1 - \delta$,

1. $x_t(a)^\top \hat{W}_t(a) \theta_t \leq x_t(a)^\top W_* \theta_t$, for all arms $a \in [K]$, for all time $t$.
2. $\| \sum_{t=1}^{T} (\hat{W}_t(a_t) - W_*)^\top x_t(a_t) \| \leq \|1_d\| \left( 2m \sqrt{T \ln ((d+td)/\delta)} \ln(T) \right).$

3.2 Algorithm and Regret Analysis

At every step $t$, the algorithm constructs optimistic estimates of the unknown weight matrix, as described in the previous section, for vector $\theta_t$ which is itself generated by an online learning subroutine. The algorithm then picks the arm which appears to be the best according to these optimistic estimates.

**Algorithm 1 Algorithm for Feasibility problem**

```
Initialize $\hat{W}_1, G_1, \text{OL-Predict with domain } \Omega = \{ \|\theta\|_* \leq 1 \}$, and $\theta_1$.

for all $t = 1 \ldots T$ do

Observe $x_t$.

Compute $\hat{W}_t(a)$ as defined in (5), for all $a$.

Play arm $a_t := \arg \min_{a \in [K]} x_t(a)^\top \hat{W}_t(a) \theta_t$.

Observe $v_t(a_t)$.

Use $x_t(a_t)$ and $v_t(a_t)$ to obtain $\hat{W}_{t+1}$ and $G_{t+1}$.

$\theta_{t+1} := \text{OL-Predict}(\theta_t, g_1, \ldots, \theta_t, g_t)$ with $g_t(\theta) = x_t(a_t)^\top \hat{W}_t(a_t) \theta - h_S(\theta)$.

end for
```
Theorem 2. Let $\mathcal{R}(T)$ be the regret for OCO with $\{g_t, \theta_t\}$, as given by Lemma 4. Specifically, $\mathcal{R}(T) = O(\sqrt{dT})$ when distance is given by $\ell_2$ norm and $\mathcal{R}(T) = O(\sqrt{\log(d)T})$ for $\ell_\infty$. Algorithm 1 achieves the following regret bound, with probability $1 - \delta$.

$$d(\frac{1}{T} \sum_t v_t(a_t), S) \leq \mathcal{R}(T)/T + O\left( m\|1_d\|\sqrt{\ln(mdT/\delta) \ln(T)/T}\right).$$

Proof. Let $H_{t-1}$ be the history of plays and observations before time $t$, i.e. $H_{t-1} := \{\theta_\tau, X_\tau, a_\tau, v_\tau(a_\tau), \tau = 1, \ldots, t-1\}$. Note that $H_{t-1}$ determines $\theta_t, \hat{W}_t, G_t$, but it does not determine $X_t, a_t, \hat{W}_t$ (since $a_t$ and $\hat{W}_t(a)$ depend on the context $X_t$ at time $t$). The proof is in 3 steps:

Step 1: Since $E[v_t(a_t)|X_t, a_t, H_{t-1}] = W^*_t X_t(a_t)$, it is sufficient to bound the distance to $S$ of the average of the vectors $W^*_t X_t(a_t)$ and then apply Azuma-Hoeffding to get the required result.

Step 2: From Corollary 2, with probability $1 - \delta$,

$$\left\| \frac{1}{T} \sum_{t=1}^T (W_t - \hat{W}_t(a_t))^\top X_t(a_t) \right\| \leq O\left( m\|1_d\|\sqrt{\ln(mdT/\delta) \ln(T)/T}\right).$$

(6)

It is therefore sufficient to bound the distance to $S$ of the average of the vectors $\hat{W}_t(a_t)^\top X_t(a_t)$.

Step 3: We use the shorthand notation of $\tilde{v}_t := \hat{W}_t(a_t)^\top X_t(a_t)$ and $\tilde{v}_{avg} := \frac{1}{T} \sum_{t=1}^T \tilde{v}_t$ for the rest of this proof. The proof is completed by showing that with probability $1 - \delta$,

$$d(\tilde{v}_{avg}, S) \leq \mathcal{R}(T)/T + \|1_d\|\sqrt{2\ln(dT/\delta)/T}.$$ 

(7)

The proof of (7) is as follows. Since we used online learning with $g_t(\theta) = \tilde{v}_t^\top \theta - h_S(\theta)$ to predict $\theta_t$, therefore, from (3) and the definition of $\mathcal{R}(T)$, we have that

$$d(\tilde{v}_{avg}, S) = \max_{\|\theta\|_* \leq 1} \{\theta^\top \tilde{v}_{avg} - h_S(\theta)\} \leq \frac{1}{T} \sum_{t=1}^T (\theta_t^\top \tilde{v}_t - h_S(\theta_t)) + \mathcal{R}(T)/T.$$ 

(8)

Now, for all $t$,

$$E_{X_t \sim D}\{\theta_t^\top \tilde{v}_t | H_{t-1}\} = E_{X_t \sim D}\{\theta_t^\top \hat{W}_t(a_t)^\top X_t(a_t) | H_{t-1}\} \leq E_{X_t \sim D, a \sim q^*(X_t)}\{\theta_t^\top \hat{W}_t(a)^\top X_t(a) | H_{t-1}\} \leq E_{X_t \sim D, a \sim q^*(X_t)}\{\theta_t^\top W^*_t X_t(a) | H_{t-1}\} \text{ (with probability } 1 - \delta) = \theta_t^\top E_{X \sim D}\{W^*_t X q^*(X)\} = \theta_t^\top v(q^*).$$

(9)

The first inequality above is due to the choice of $a_t$ in the algorithm. The second inequality is from the first part of Corollary 2. This gives $E[\theta_t^\top \tilde{v}_t | H_{t-1}] - h_S(\theta_t) \leq \theta_t^\top v(q^*) - h_S(\theta_t) \leq d(v(q^*), S) = 0$. The proof is now completed by using Azuma-Hoeffding to bound $\sum_{t=1}^T \theta_t^\top \tilde{v}_t - h_S(\theta_t) \leq \|1_d\|\sqrt{2\ln(dT/\delta)}$, with probability $1 - \delta$, and using this in (8).

4 The general case (linCBwCR)

In this section, we consider the linCBwCR problem (Definition 1). If we know the value of OPT := $f(v(q^*))$, we can reduce this problem to the Feasibility Problem (linCBwC) by defining the constraint set to be

$$S' = \{v : f(v) \geq \text{OPT}, v \in S\}.$$
A common approach \cite{20,7} to eliminate the assumption that OPT is known is to get increasingly accurate estimates of OPT computed at geometric time intervals as the algorithm progresses. However this approach does not seem to be directly employable here. As we shall see later, due to bandit feedback in this problem, estimating OPT seems to require “pure exploration” for some rounds where the arms picked are tailored to getting such an estimate. Further, the error in the estimate thus obtained using $T_0$ rounds turns out to be of form $\hat{O}(m\|1_d\|Z^*/\sqrt{T_0})$, where $Z^*$ is a parameter (defined later) that captures the tradeoff between the objective and constraints. Thus, the knowledge or estimation of $Z^*$ is required for this approach in addition to estimating OPT. Instead, following \cite{5}, we use an approach that requires only the knowledge or even an $O(1)$ approximation of $Z^*$.

Let OPT$^\gamma$ denote the value of the optimal policy when the feasibility constraint is relaxed from $v(q) \in S$ to $d(v(q), S) \leq \gamma$ for any $\gamma \geq 0$.

$$\text{OPT}^\gamma := \max_q f(v(q)) \text{ such that } d(v(q), S) \leq \gamma$$  \hspace{1cm} (10)

Lemma 5. There exists $Z \geq 0$ such that for all $\gamma \geq 0$, $\text{OPT}^\gamma \leq \text{OPT} + \frac{Z}{2} \gamma$. Let $Z^*$ denote the minimum value of such $Z$.

Assumption 1. Assume we are given $Z \geq L$ satisfying the condition in Lemma 5.

Below, we present an algorithm under Assumption 1. The algorithm is similar to that in the previous section: it constructs estimates of the weight matrix exactly as in Section 3.1. Here, we run two instances of the online learning algorithm: we search over the domain of Fenchel dual variables for the objective in addition to that for the constraints as in the previous section. The optimistic estimates are constructed using the vector obtained on combining the dual variables for the objective and those for the constraints using $Z$. The algorithm then chooses the arm which appears to be the best according to optimistic estimates, as before.

Algorithm 2 Algorithm for linCBwCR

Initialize $\tilde{W}_1, \tilde{G}_1$, two instances of OL-Predict with domain $\Omega_1 = \{\theta : \|\theta\|_* \leq 1\}$, and $\Omega_2 = \{\phi : \|\phi\|_* \leq L\}$, $\theta_1$ and $\phi_1$.

for all $t = 1, \ldots, T$ do

Observe $X_t$.

For every $a \in [K]$, compute $\tilde{W}_t(a) := \arg\max_{W \in \tilde{G}_t} x_t(a)^T W (\phi_t + Z \theta_t)$.

Play the arm $a_t := \arg\min_{a \in [K]} x_t(a)^T \tilde{W}_t(a) (\theta_t + Z \theta_t)$.

Observe $v_t(a_t)$.

Use $x_t(a_t)$ and $v_t(a_t)$ to obtain $\tilde{W}_{t+1}$ and $\tilde{G}_{t+1}$.

$\theta_{t+1} := \text{OL-Predict}_1(\theta_1, g_1, \ldots, \theta_t, g_t)$ with $g_t(\theta) = x_t(a_t)^T \tilde{W}_t(a_t) \theta - h_S(\theta)$.

$\phi_{t+1} := \text{OL-Predict}_2(\phi_1, \psi_1, \ldots, \phi_t, \psi_t)$ with $\psi_t(\theta) = x_t(a_t)^T \tilde{W}_t(a_t) \phi - f^*(\phi)$.

end for

Theorem 3. Given a $Z \geq L$ as per Assumption 1, Algorithm 2 achieves the following bounds with probability $1 - \delta$:

$$\text{avg-regret}_1(T) \leq Z \cdot O \left( \frac{R(T)}{T} + m\|1_d\| \sqrt{\ln(dT/\delta)\ln(T)/T} \right),$$

$$\text{avg-regret}_2(T) \leq O \left( \frac{R(T)}{T} + m\|1_d\| \sqrt{\ln(dT/\delta)\ln(T)/T} \right).$$

Here $R(T) = R_1(T) + \frac{R_2(T)}{m}$, with $R_1(T)$ being the regret for OCO with $\{g_t, \theta_t\}$, and $R_2(T)$ being the regret for OCO with $\{\psi_t, \phi_t\}$, as given by Lemma 4. Specifically, $R(T) = O(\sqrt{T})$ when $\ell_2$ norm is used and $R(T) = O(\sqrt{\log(dT)})$ for $\ell_\infty$ norm.

\textsuperscript{2}As an application of this approach, for the special case of budget constraints (linCBwK), we show how to get an approximation of $Z^*$ using the initial few rounds, giving a sub-linear regret overall; see Section 5 for details.
5 Budget Constraints (linCBwK)

For the special case of linCBwK, knowing or estimating \( \text{OPT} \) is sufficient to estimate \( Z \).

**Lemma 6.** \( \frac{Z}{2} = \frac{\text{OPT}}{(B/T)} \) satisfies the condition in Lemma 5, when \( S \) is given by budget constraints and \( \ell_\infty \) norm is used for distance.

Therefore, if we know \( \text{OPT} \), we can use Algorithm 2 with above value of \( Z \) plus 1, and distance defined by \( \ell_\infty \) norm, to achieve an \( \tilde{O}(\left(\frac{\text{OPT}}{(B/T)} + 1\right) \frac{m}{\sqrt{T}}) \) regret in objective (as given by Theorem 3). We show how to estimate \( Z \) as part of the algorithm for the special case of linCBwK, using the first few rounds. Also, we are not allowed to violate the budgets at all for this problem, so the algorithm needs to be modified to handle this. Given parameters \( T_0 \) and \( B' \), the algorithm is as follows.

In the first \( T_0 \) rounds do a “pure exploration” and estimate a \( Z \) satisfying Lemma 5. Set aside a \( B' \) amount from the remaining budget for each dimension. Run Algorithm 2 for the remaining time steps with the remaining budget. Stop if the full actual budget \( B \) is consumed for any dimension.

**Theorem 4.** Using the above algorithm with \( T_0 \) such that \( B > \max\{T_0, mT/\sqrt{T_0}\} \), \( B' = m\sqrt{T\ln(dT/\delta)\ln(T)} \) and an appropriate estimation procedure for \( Z \), we get a high probability regret bound of

\[
\tilde{O}\left(\left(\frac{\text{OPT}}{(B/T)} + 1\right) \frac{T_0}{T} + \frac{m}{\sqrt{T}}\right).
\]

In particular, using \( T_0 = \sqrt{T} \), and assuming \( B > mT^{3/4} \) gives a regret bound of

\[
\tilde{O}\left(\left(\frac{\text{OPT}}{(B/T)} + 1\right) \frac{m}{\sqrt{T}}\right).
\]

The proof of Theorem 4 uses the following estimate for \( Z \) obtained as described in Section 5.1.

**Lemma 7.** Let \( \gamma = 2m\sqrt{\ln(T_0)\ln(T_0d/\delta)/T_0} \). Using the first \( T_0 \) rounds for exploration, we can obtain \( Z \) such that with probability \( 1 - O(\delta) \),

\[
\frac{\text{OPT}}{(B/T)} + 1 \leq Z \leq (1 + \frac{7\gamma}{(B/T)})(\frac{\text{OPT}}{(B/T)} + 1).
\]

Lemma 7 implies that as long as \( \frac{B}{T} \geq \gamma \), i.e., \( B \geq \Omega\left(\frac{mT}{\sqrt{T_0}}\right) \), \( Z \) is a constant factor approximation of \( \frac{\text{OPT}}{(B/T)} + 1 \geq Z^* \), therefore Theorem 3 provides an \( \tilde{O}\left(\left(\frac{\text{OPT}}{(B/T)} + 1\right) \frac{m}{\sqrt{T}}\right) \) regret bound if the algorithm is not stopped due to budget violation. As implied by the bound on avg-regret\(^2(T) \) in Theorem 3, setting aside \( B' \) budget ensures that budget is not violated with high probability. However, the reduction of the budget by \( B' \) incurs additional \( \frac{B'}{B} \)OPT regret. Also, using the first \( T_0 \) rounds for pure exploration can consume (at most) \( T_0 \) amount from the budget, incurring an additional regret of \( \frac{T_0}{B} \)OPT. Together, these observations give Theorem 4.

5.1 Estimating OPT

Here we sketch the high level idea behind obtaining the result in Lemma 7 by estimating \( \text{OPT} \). Due to bandit feedback, estimating \( \text{OPT} \) in our problem has significant challenges compared to the full information settings like those in [5, 20, 7]. The approach used in those papers is to use the value of the policy that performs the best on the observations seen so far. However, in case of bandit feedback, observations are made only for the arms picked by the algorithm, therefore, it is not possible to directly evaluate any policy other than the one that was followed in the history. A fix for this, inspired by [22, 3], is to play every arm with some minimum probability in every round, so that using importance sampling we can re-weight the observations and construct an unbiased estimator of the outcome vector.
for any arm. A crude version of this that plays every arm with probability $1/K$ in every round for $T_0$ rounds would obtain an estimation error of $\tilde{O}(KmZ^*/\sqrt{T_0})$. Instead, we provide a method that takes advantage of the linear structure of the problem, and explores in the $m$-dimensional space of contexts and weight vectors to obtain bounds independent of $K$.

We use the following procedure. In every round $t = 1, \ldots, T_0$, after observing $X_t$, let $p_t \in \Delta^K$ be

$$p_t := \arg \max_{p \in \Delta^K} \|X_tp\|_{M_t^{-1}}^{-1},$$  \hspace{1cm} (11)

where $M_t := I + \sum_{i=1}^{t-1} (X_ip_i)(X_ip_i)^\top$.  \hspace{1cm} (12)

Select arm $a_t = a$ with probability $p_t(a)$. In fact, since $M_t$ is a PSD matrix, due to convexity of the function $\|Xtp\|^2_{M_t^{-1}}$, it is the same as playing $a_t = \arg \max_{a \in [K]} \|X_t(a)\|_{M_t^{-1}}$.

Construct an estimate $\hat{W}_t$ of $W_*$ at time $t$ as

$$\hat{W}_t := M_t^{-1} \sum_{i=1}^{t-1} (X_ip_i)\mathbf{v}_i(a_i)^\top.$$  

And, for some value of $\gamma$ defined later, obtain an estimate $\hat{OPT}^\gamma$ of OPT as:

$$\hat{OPT}^\gamma := \max_{q} \frac{f(\frac{1}{T_0} \sum_{i=1}^{T_0} \hat{W}_iq(X_t))}{d(\frac{1}{T_0} \sum_{i=1}^{T_0} \hat{W}_iq(X_t)), S) \leq \gamma}.  \hspace{1cm} (13)$$

For an intuition about the choice of arm in (11), observe from the discussion in Section 2.1 that every column $w_{xj}$ of $W_*$ is guaranteed to lie inside the confidence ellipsoid centered at column $\tilde{w}_{tj}$, namely the ellipsoid, $\|\mu - \tilde{w}_{tj}\|^2_{M_t} \leq 4m \ln(Tm/\delta)$. Note that this ellipsoid has principle axes as eigenvectors of $M_t$, and the length of semi-principle axes is given by inverse eigenvalues of $M_t$.

Therefore, by maximizing $\|Xtp\|^2_{M_t^{-1}}$, we are choosing the context closest to the direction of the longest principal axes of the confidence ellipsoid, i.e. in the direction of maximum uncertainty. Intuitively, this corresponds to pure exploration: by making an observation in the direction where uncertainty is large we can reduce the uncertainty in our estimate most effectively.

A more algebraic explanation is as follows. For a good estimation of OPT by $\hat{OPT}^\gamma$, we want the estimates $\hat{W}_t$ and $W_*$ to be close enough so that $\frac{1}{T_0} \sum_{i=1}^{T_0} (\hat{W}_t - \hat{W}_a)^\top X_tq(X_t)$ is small for all policies $q$, and in particular for optimal policies. Now, using Cauchy-Schwartz this is bounded by

$$\frac{1}{T_0} \sum_{i=1}^{T_0} \|1_d\| \|\hat{W}_t - W_*\|_{M_t} \|X_tq(X_t)\|_{M_t^{-1}},$$

where we define $\|W\|_{M_t}$, the $M$-norm of matrix $W$, to be the max of column-wise $M$-norms. Using Lemma 1, the term $\|\hat{W}_t - W_*\|_{M_t}$ is bounded by $2\sqrt{m \ln(T_0md/\delta)}$ with probability $1 - \delta$. Lemma 2 bounds the second term $\sum_{i=1}^{T_0} \|X_tq(X_t)\|_{M_t^{-1}}$, but only when $q$ is the played policy. This is where we use that the played policy $p_t$ was chosen to maximize $\|X_tp\|_{M_t^{-1}}$, so that $\sum_{i=1}^{T_0} \|X_tq(X_t)\|_{M_t^{-1}} \leq \sum_{i=1}^{T_0} \|X_tq(X_t)\|_{M_t^{-1}}$ and the bound $\sum_{i=1}^{T_0} \|X_tq(X_t)\|_{M_t^{-1}} \leq \sqrt{mT_0 \ln(T_0)}/(T_0)$ given by Lemma 2 actually bounds $\sum_{t=1}^{T_0} \|X_tq(X_t)\|_{M_t^{-1}}$ for all $q$.

We prove the following lemma.

**Lemma 8.** For $\gamma = 2\|1_d\|m \sqrt{\ln(T_0) \ln(T_0d/\delta)}/T_0$, with probability $1 - O(\delta)$, $OPT - L\gamma \leq \hat{OPT}^{2\gamma} \leq OPT + 6\gamma(Z^* + L)$.

For linCBwK, $Z^* \leq \frac{OPT}{(B/T)}$, $L = 1$, $\|1_d\| = \|1_d\|_\infty = 1$. Therefore, $Z = (\frac{OPT^{2\gamma} + L}{(B/T)}) + 1$ satisfies the bounds of Lemma 7.
References


Appendix

A Concentration Inequalities

Lemma 9 (Azuma-Hoeffding inequality). If a super-martingale \((Y_t; t \geq 0)\), corresponding to filtration \(\mathcal{F}_t\), satisfies \(|Y_t - Y_{t-1}| \leq c_t\) for some constant \(c_t\), for all \(t = 1, \ldots, T\), then for any \(a \geq 0\),

\[
\Pr(Y_T - Y_0 \geq a) \leq e^{-\frac{a^2}{2 \sum_{t=1}^{T} c_t^2}}.
\]

B Confidence ellipsoids

Proof of Corollary 1. The following holds with probability \(1 - \delta\).

\[
\sum_{t=1}^{T} |\bar{\mu}_t^T x_t - \mu_*^T x_t| \leq \sum_{t=1}^{T} \|\bar{\mu}_t - \mu_*\|_{M_t} \|x_t\|_{M_t^{-1}}
\]

\[
\leq \left( \sqrt{m \ln \left( \frac{1 + tm}{\delta} \right)} + \sqrt{m} \right) \sqrt{mT \ln(T)}.
\]

The inequality in the first line is a matrix-norm version of Cauchy-Schwartz (Lemma 10). The inequality in the second line is due to Lemmas 1 and 2. The lemma follows from multiplying out the two factors in the second line.

\[]

Lemma 10. For any positive definite matrix \(M \in \mathbb{R}^{n \times n}\) and any two vectors \(a, b \in \mathbb{R}^n\), \(|a^T b| \leq \|a\|_M \|b\|_{M^{-1}}\).

Proof. Since \(M\) is positive definite, there exists a matrix \(M_{1/2}\) such that \(M = M_{1/2} M_{1/2}^T\). Further, \(M^{-1} = M_{-1/2}^T M_{-1/2}\) where \(M_{-1/2} = M_{1/2}^{-1}\).

\[
\|a^T M_{1/2}\|^2 = a^T M_{1/2} M_{1/2}^T a = a^T M a = \|a\|_M^2.
\]

Similarly, \(|M_{-1/2} b|^2 = \|b\|_{M^{-1}}^2\). Now applying Cauchy-Schwartz, we get that

\[
|a^T b| = |a^T M_{1/2} M_{-1/2} b| \leq \|a^T M_{1/2}\| \|M_{-1/2} b\| = \|a\|_M \|b\|_{M^{-1}}.
\]

\[]

Proof of Corollary 2. Here, the first claim follows simply from definition of \(\hat{W}_t(a)\) and the observation that with probability \(1 - \delta\), \(W^* \in \mathcal{G}_t\). To obtain the second claim, apply Corollary 1 with \(\mu_* = w_{*j}, y_t = x_t(a_t), \bar{\mu}_t = [\hat{W}_t(a_t)]_j\) (the \(j^{th}\) column of \(\hat{W}_t(a_t)\)), to bound \(|\sum_t ([\hat{W}_t(a_t)]_j - w_{*j})^T x_t(a_t)| \leq \sum_t |([\hat{W}_t(a_t)]_j - w_{*j})^T x_t(a_t)|\) for every \(j\), and then take the norm.

\[]

C Appendix for Section 4 (linCBwCR)

Proof of Lemma 5. The proof of this lemma is similar to the proof of Lemma 5.1 in [5]. Let \(\Omega := \{y : \exists q, y = \mathbb{E}_{X \sim \mathcal{D}}[W^T X q(X)]\}\). Then, \(\Omega\) is a convex set. And, \(\text{OPT}^\gamma\) can be written as the following convex optimization problem over \(y\)

\[
\text{OPT}^\gamma := \max_{y \in \Omega} f(y) \text{ such that } d(y, S) \leq \gamma.
\]
Then, applying Lagrangian duality for convex programs
\[ \text{OPT}^\gamma = \min_{\lambda \geq 0} \max_{y \in \Omega} f(y) + \lambda (\gamma - d(y, S)) \]

By Fenchel duality, for any L-Lipschitz concave function \( f \),
\[ f(y) = \min_{\phi : \|\phi\| \leq L} f^*(\phi) - \phi^T y, \]
where \( f^* \) is the Fenchel dual of \( f \). Specifically, for distance function,
\[ d(y, S) = \max_{\theta : \|\theta\| \leq 1} \theta^T y - h_S(\theta), \]
where for any convex set \( S \), \( h_S \) is defined as: \( h_S(\theta) := \max_{v \in S} \theta^T v \). Substituting in the expression for \( \text{OPT}^\gamma \), we get
\[ \text{OPT}^\gamma = \min_{\lambda \geq 0} \max_{y \in \Omega} \min_{\|\phi\| \leq L, \|\theta\| \leq 1} (f^*(\phi) - \phi^T y + \lambda (\gamma - \theta^T y - h_S(\theta))) \]
\[ = \min_{\lambda \geq 0} \max_{\|\phi\| \leq L, \|\theta\| \leq 1} \min_{y \in \Omega} f^*(\phi) - \phi^T y - \lambda \theta^T y + \lambda h_S(\theta) + \lambda \gamma \]
\[ = \min_{\lambda \geq 0} \max_{\|\phi\| \leq L, \|\theta\| \leq 1} f^*(\phi) + h_\Omega(-\phi - \lambda \theta) + \lambda h_S(\theta) + \lambda \gamma \]
\[ = \text{OPT} + \lambda^* \gamma \]

Here, \( \phi^*, \theta^*, \lambda^* \) denote the optimal value of these variables for the corresponding optimization problem for \( \text{OPT} \). Therefore, \( Z = \lambda^* \) satisfies the condition in Lemma 5. \( \square \)

**Proof of Theorem 3.** The proof is an extension of proof of Theorem 2, and uses the properties of \( Z \) to handle the objective in addition to constraints.

We use the shorthand notation of \( \tilde{v}_t := \tilde{W}_t(a_t)^T x_t(a_t) \), \( \tilde{v}_{\text{avg}} = \frac{1}{T} \sum_{t=1}^T \tilde{v}_t \) in this section. Similar to the observation in Equation (8) in the proof of Theorem 2, using online learning guarantees we can obtain that for the choice of \( \theta_t, \phi_t \) made by this algorithm,
\[ d(\tilde{v}_{\text{avg}}, S) \leq \frac{1}{T} \sum_{t=1}^T \left( \theta_t^T \tilde{v}_t - h_S(\theta_t) \right) + \frac{R_1(T)}{T}. \] (14)
\[ f(\tilde{v}_{\text{avg}}) \geq \frac{1}{T} \sum_{t=1}^T \left( f^*(\phi_t) - \phi_t^T \tilde{v}_t \right) + \frac{R_2(T)}{T}. \] (15)

For all \( t \), since the algorithm selected \( a_t \in [K] \) to minimize \( x_t(a)^T \tilde{W}_t(a)(\phi_t + Z \theta_t) \), and because \( \tilde{W}_t(a) \) is an optimistic estimate for every \( a \), therefore using similar line of arguments as in Equation (9), we get that with probability \( 1 - \delta \),
\[ \mathbb{E}[f(\phi_t) + Z h_S(\theta_t) - (\phi_t + Z \theta_t)^T \tilde{v}_t | H_{t-1}] \geq f^*(\phi_t) + Z h_S(\theta_t) - (\phi_t + Z \theta_t)^T v(q^*). \] (16)

Below are the details of this argument.
\[ \mathbb{E}[(\phi_t + Z \theta_t)^T \tilde{v}_t | H_{t-1}] = \mathbb{E}[(\phi_t + Z \theta_t)^T \tilde{W}_t(a_t)^T x_t(a_t) | H_{t-1}] \]
\[ \leq \mathbb{E}[(\phi_t + Z \theta_t)^T \left( \sum_a \tilde{W}_t(a)^T x_t(a)q^*(X_t, a) \right) | H_{t-1}] \]
\[ \leq \mathbb{E}[(\phi_t + Z \theta_t)^T W_s^T X_t q^*(X_t) | H_{t-1}] \text{ (with probability } 1 - \delta) \]
\[ = (\phi_t + Z \theta_t)^T \mathbb{E}_{X \sim \mathcal{D}} \left[ W_s^T X q^*(X) \right] = (\phi_t + Z \theta_t)^T v(q^*). \]
Now, the term on the right hand side in Equation (16) is greater than or equal to

\[
\left( \min_{\|\phi\| \leq L} f^*(\phi_t) - \phi^\top v(q^*) \right) - Z \left( \max_{\|\theta\| \leq 1} \theta^\top v(q^*) - h_S(\theta_t) \right) = f(v(q^*)) - Z d(v(q^*),S) = \text{OPT}
\]

where the second last equality uses Fenchel duality, and the last equality uses \(d(v(q^*),S) = 0\). Therefore, for every \(t\), the conditional expectation on the left hand side of (16) is lower bounded by OPT. Summing over \(t\), and using Azuma-Hoeffding, we obtain that probability \(1 - \delta\),

\[
\sum_{t=1}^T f^*(\phi_t) - \phi^\top v_t - Z(\theta_t^\top v_t - h_S(\theta_t)) \geq T\text{OPT} - \|1_d\|\sqrt{2T\ln(dT/\delta)}
\]

Combining Equation (14) and (15) using \(Z\), and substituting the bound from (17), we get

\[
f(\bar{v}_{\text{avg}}) - Zd(\bar{v}_{\text{avg}},S) \geq \text{OPT} - \|1_d\|\sqrt{2T\ln(dT/\delta)} - Z\frac{R(T)}{T}.
\]

Now, Step 1 and Step 2 in the proof of Theorem 2 hold as it is, so that \(\bar{v}_{\text{avg}} := \frac{1}{T} \sum_{t=1}^T v_t\) is close to \(v_{\text{avg}} := \frac{1}{T} \sum_{t=1}^T v_t(a_t)\). Therefore, we can replace \(\bar{v}_{\text{avg}}\) by \(v_{\text{avg}}\) in above to get with probability \(1 - \delta\)

\[
f(v_{\text{avg}}) - Zd(v_{\text{avg}},S) \geq \text{OPT} - Z O \left( \frac{m\|1_d\|}{T} \sqrt{T\ln(mdT/\delta)} \ln(T) + \frac{R(T)}{T} \right).
\]

Since \(d(v_{\text{avg}},S) \geq 0\), this proves the bound on \(\text{avg-regret}^1(T)\).

To obtain the bound on distance from the constraint set, first note that

\[
\mathbb{E}[f(v_{\text{avg}})] \leq \text{OPT}\mathbb{E}[d(v_{\text{avg}},S)]
\]

To see this, let the random variable \(\bar{q}\) denote the empirical played policy, i.e. the policy implied by the contexts observed and the arms played \(X_1, a_1, \ldots, X_T, a_T\), so that by definition \(v(\mathbb{E}q) = \mathbb{E}[v_{\text{avg}}]\), and \(d(v(\mathbb{E}q),S) = d(\mathbb{E}[v_{\text{avg}}],S) \leq \mathbb{E}[d(v_{\text{avg}},S)]\). Therefore, \(\mathbb{E}\bar{q}\) is a feasible solution for (10) with \(\gamma = \mathbb{E}[d(v_{\text{avg}},S)]\), implying \(f(v(\mathbb{E}\bar{q})) \leq \text{OPT}\gamma\). Also, by concavity of \(f\) and definition of policy \(\bar{q}\), \(\mathbb{E}[f(v_{\text{avg}})] \leq f(\mathbb{E}[v_{\text{avg}}]) = f(v(\mathbb{E}\bar{q}))\). Therefore, the above observation follows.

Now, using the assumption about the parameter \(Z\),

\[
\mathbb{E}[f(v_{\text{avg}})] - \frac{Z}{2}\mathbb{E}[d(v_{\text{avg}},S)] \leq \text{OPT}\mathbb{E}[d(v_{\text{avg}},S)] - \frac{Z}{2}\mathbb{E}[d(v_{\text{avg}},S)] \leq \text{OPT}
\]

Using (18), this implies

\[
\mathbb{E}[f(v_{\text{avg}})] - \frac{Z}{2}\mathbb{E}[d(v_{\text{avg}},S)] \leq \text{OPT} \leq \mathbb{E}[f(v_{\text{avg}})] - Z\mathbb{E}[d(v_{\text{avg}},S)] + Z O \left( \frac{m\|1_d\|}{T} \sqrt{T\ln(mdT/\delta)} \ln(T) + \frac{R(T)}{T} \right).
\]

Rearranging the terms, we get

\[
\frac{Z}{2}\mathbb{E}[d(v_{\text{avg}},S)] \leq Z O \left( \frac{m\|1_d\|}{T} \sqrt{T\ln(mdT/\delta)} \ln(T) + \frac{R(T)}{T} \right).
\]

Then, using \(d(\mathbb{E}[v_{\text{avg}}],S) \leq \mathbb{E}[d(v_{\text{avg}},S)]\) and applying Azuma-Hoeffding, we get the desired bound on \(\text{avg-regret}^2(T)\) with high probability. \(\square\)
D Appendix for Section 5 (linCBwK)

Proof of Lemma 6. The proof of this lemma is similar to the proof of Lemma 4 in [6]. In linCBwK problem, \( \text{OPT}^\gamma \) represents the value of optimal policy when the budget constraints are relaxed by \( \gamma \), i.e. for feasibility set \( S^\gamma = \{ v : v \leq \frac{\delta}{\gamma}1 + \gamma \} \). Now, suppose for contradiction that \( \text{OPT}^\gamma > \text{OPT} + \frac{\text{OPT}^\gamma}{B/T} \). Let \( \hat{q} \) be the optimal policy for \( \text{OPT}^\gamma \). Then, one could construct a feasible policy \( q' \) for the original problem with budget \( B \) by scaling down the probabilities given by \( \hat{q} \) by \( \frac{B}{B+\gamma T} \). (And do nothing with the remaining probability). Since \( \hat{q} \) consumed at most \( B + \gamma T \) budget, \( q' \) will consume at most \( B \) budget. Also, the value of policy \( q' \) will be at least \( \text{OPT}^\gamma \frac{B}{B+\gamma T} > (\text{OPT} + \frac{\text{OPT}^\gamma}{B/T} \gamma) \frac{B}{B+\gamma T} = \text{OPT} \). Thus we arrive at a contradiction.

\[ \square \]

E Appendix for Section 5.1 (Estimating OPT)

Proof of Lemma 8. Let us define an intermediate "sample optimal" as:

\[
\overline{\text{OPT}}^\gamma := \max_q \text{ such that } f\left(\frac{1}{T_0} \sum_{i=1}^{T_0} W^*_i X_i q(X_i)\right) \leq \gamma
\]

Above sample optimal knows the weight matrix \( W^* \), the error comes only from approximating the expected value over context distribution by average over the observed contexts. We do not actually compute \( \overline{\text{OPT}}^\gamma \), but will use it for the convenience of proof exposition. The proof involves two steps.

Step 1: Bound \( |\overline{\text{OPT}}^\gamma - \text{OPT}| \).

Step 2: Bound \( |\hat{\text{OPT}}^{2\gamma} - \overline{\text{OPT}}^\gamma| \) by bounding \( \| \sum_{i=1}^{T_0}(\hat{W}_i - W^*_i)^\top X_i q(X_i) \| \) for all \( q \).

Step 1 bound can be borrowed from the work on Online Stochastic Convex Programming in [5]: since \( W^* \) is known so there is effectively full information before making the decision, i.e., consider the vectors \( W^*_i x_t(a) \) as reward vectors which can be observed for all arms \( a \) before choosing the distribution over arms to be played at time \( t \), therefore, the setting in [5] applies. In fact, \( \overline{\text{OPT}}^\gamma \) as defined by Equation (F.10) in [5] when \( \mathcal{A}_t = \{W^*_i x_t(a), a \in [K]\} \), is same as the \( \overline{\text{OPT}}^\gamma \) defined here. And using Lemma F.4 and Lemma F.6 in [5], we obtain that for any \( \gamma \geq 2\|1_d\| m \sqrt{\ln(T_0) \ln(T_0 d/\delta)/T_0} \), with probability \( 1 - O(\delta) \),

\[
\text{OPT} - L \gamma \leq \overline{\text{OPT}}^\gamma \leq \text{OPT} + 2\gamma(Z^* + L).
\]

For Step 2, we show that with probability \( 1 - \delta \), for all \( q \), \( \gamma \geq 2\|1_d\| m \sqrt{\ln(T_0) \ln(T_0 d/\delta)/T_0} \),

\[
\frac{1}{T_0} \sum_{i=1}^{T_0}(\hat{W}_i - W^*_i)^\top X_i q(X_i) \| \leq \gamma
\]

This is sufficient to prove both lower and upper bound on \( \hat{\text{OPT}}^{2\gamma} \) for \( \gamma \geq 2\|1_d\| m \sqrt{\ln(T_0) \ln(T_0 d/\delta)/T_0} \). For lower bound, we can simply use (21) for optimal policy for \( \overline{\text{OPT}}^\gamma \), denoted by \( \hat{q} \). This implies that (because of relaxation of distance constraint by \( \gamma \)) \( \hat{q} \) is a feasible primal solution for \( \hat{\text{OPT}}^{2\gamma} \), and therefore using (20),

\[
\hat{\text{OPT}}^{2\gamma} \geq \overline{\text{OPT}}^\gamma \geq \text{OPT} - L \gamma.
\]

For the upper bound, we can use (21) for the optimal policy \( \hat{q} \) for \( \hat{\text{OPT}}^{2\gamma} \). Then, using (20),

\[
\hat{\text{OPT}}^{2\gamma} \leq \overline{\text{OPT}}^{2\gamma} \leq \text{OPT} + 6\gamma(Z^* + L).
\]
Combining, this proves the desired lemma statement:

\[
\text{OPT} - L\gamma \leq \hat{\text{OPT}}^{2\gamma} \leq \text{OPT} + 6\gamma(Z^* + L)
\]  

(22)

What remains is to proof the claim in (21). Observe that for any \( q \),

\[
\left\| \sum_{t=1}^{T_0} (\hat{W}_t - W_*)^\top X_t q(X_t) \right\| \leq \sum_{t=1}^{T_0} \left\| (\hat{W}_t - W_*)^\top X_t q(X_t) \right\|
\]

\[
\leq \sum_{t=1}^{T_0} \left\| \mathbf{1}_d \right\| \left\| \hat{W}_t - W_* \right\|_M \left\| X_t q(X_t) \right\|_{M^{-1}}
\]

where \( \left\| \hat{W}_t - W_* \right\|_M = \max_j \left\| \hat{w}_{tj} - w_{*j} \right\|_M \).

Now, applying Lemma 1 to every column \( \hat{w}_{tj} \) of \( \hat{W}_t \), we have that with probability \( 1 - \delta \) for all \( t \),

\[
\left\| \hat{W}_t - W_* \right\|_M \leq 2\sqrt{m \log(td/\delta)} \leq 2\sqrt{m \log(T_0d/\delta)}
\]

And, by choice of \( p_t \)

\[
\left\| X_t q(X_t) \right\|_{M^{-1}} \leq \left\| X_t p_t \right\|_{M^{-1}}.
\]

Also, by Lemma 2,

\[
\sum_{t=1}^{T_0} \left\| X_t p_t \right\|_{M^{-1}} \leq \sqrt{mT_0 \ln(T_0)}
\]

Therefore, substituting,

\[
\left\| \sum_{t=1}^{T_0} (\hat{W}_t - W_*)^\top X_t q(X_t) \right\| \leq \left\| \mathbf{1}_d \right\| (2\sqrt{m \log(T_0d/\delta)}) \sum_{t=1}^{T_0} \left\| X_t p_t \right\|_{M^{-1}}
\]

\[
\leq \left\| \mathbf{1}_d \right\| (2\sqrt{m \log(T_0d/\delta)}) \sqrt{mT_0 \ln(T_0)}
\]

\[
\leq T\gamma
\]

\( \Box \)