Latent Contextual Bandits: A Non-Negative Matrix Factorization Approach

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sanjay Shakkottai

The University of Texas at Austin

June 7, 2016

Abstract

We consider the stochastic contextual bandit problem with a large number of observed contexts and arms, but with a latent low-dimensional structure across contexts. This low dimensional (latent) structure encodes the fact that both the observed contexts and the mean rewards from the arms are convex mixtures of a small number of underlying latent contexts. At each time, we are presented with an observed context; the bandit problem is to determine the corresponding arm to pull in order to minimize regret. Assuming a separable and low rank latent context vs. mean-reward matrix, we employ non-negative matrix factorization (NMF) techniques on sub-sampled estimates of matrix entries (estimates constructed from careful arm sampling) to efficiently discover the underlying factors. This estimation lies at the core of our proposed \( \epsilon \)-greedy NMF-Bandit algorithm that switches between arm exploration to reconstruct the reward matrix, and exploitation of arms using the reconstructed matrix in order to minimize regret. We identify singular value conditions on the non-negative factors under which the NMF-Bandit algorithm has \( O(L\text{poly}(m, \log K) \log T) \) regret where \( L \) is the number of observed contexts, \( K \) is the number of arms, and \( m \) is the number of latent contexts. We further propose a class of generative models that satisfy our sufficient conditions, and derive a lower bound that matches our achievable bounds up to a \( \text{poly}(m, \log K) \) factor. Finally, we validate the NMF-bandit algorithm on synthetic data-sets.

1 Introduction

The study of bandit problems captures the inherent tradeoff between exploration and exploitation in online decision making. In various real world settings, policy designers have the freedom of observing specific samples and learning a model of the collected data on the fly; this online learning is instrumental in making future decisions. For instance in movie recommendations, algorithms suggest movies to users in order to simultaneously satisfy them and learn their preferences in an online manner. Similarly, for product recommendations (e.g.
in Amazon) or web advertisement, there is an inherent tradeoff between collection of training
data for user preferences, and recommending the best items that maximize profit according
to the currently learned model. Multi-armed bandit problems provide a principled approach
to attain this delicate balance between exploration and exploitation [6].

The fundamental $K$-armed bandit problem has been studied extensively for several decades.
In the stochastic setting, one is faced with the choice of pulling one arm during each time-slot
among $K$ arms, where the $k^{th}$ arm has mean reward $U_k$. The task is to accumulate a total
reward as close as possible to a genie strategy that has prior knowledge of arm statistics,
and always selects the optimal arm in each time-slot. The expected difference between the
rewards collected by the genie strategy and the online strategy is defined as the regret. The
expected regret of the state of the art algorithms [6] scales as $O(K \log t)$.

In contextual bandits, the algorithm has access to side-information called observed contexts
$s \in \{1, 2 \ldots L\}$ – these observed contexts influence the reward statistics of the $K$ arms. Here,
there are $(K \times L)$ reward parameters $U_{sk}$ to learn: for each arm and for each observed context.
As an example, the past browsing history or demographic information can form the observed
contexts, and these contexts influence the expected rewards of various items in an online
storefront. Since there are $(K \times L)$ reward parameters, it has been shown [6, 30] that the
best expected regret obtainable scales as $O(KL \log t)$ compared to a genie who knows all the
reward parameters $U_{sk}$.

In many practical settings that are encountered today, the number of observed contexts
can be very large. For instance, in an online storefront such as Amazon, the observed contexts
would include the user identity, demographics, purchase history, friends’ purchases, and social
information if available. Instead of dealing with every possible side-information combination
as a distinct observed context, low-rank models can dramatically reduce such contextual
information into a small dimension that is much easier to learn and exploit. For example,
the Netflix data-set (ratings matrix of movies vs. users) which has more than 100 million
movie ratings can be approximated surprisingly well by matrices of rank as low as 20 [5].
This means that only 20 latent dimensions suffice to predict user preferences, whereas the
observed context would include the user identity and their entire past watching history (much
higher dimension).

In this paper we develop a principled approach to combine offline algorithms for learning
latent dimensions with intelligent online sampling techniques. Specifically, we aim to capture
the online exploration and exploitation tradeoff while learning low-dimensional structure from
real data. We consider a variation of the contextual bandit setting, where the reward vectors
corresponding to a given context is a convex combination of the rewards of a few $(m)$ latent contexts
where $m << L, K$.

### 1.1 Modeling Latent Contextual Bandits

We model the composite mean reward matrix to be an $(L \times K)$ matrix $U = AW$. In other
words, $U_{lk}$ is the expected reward if arm $k$ is pulled when the observed context is $l$. The
matrix $A$ is a $(L \times m)$ matrix that encodes the relation between the $L$ observed contexts
and the $m$ latent contexts, with each row of $A$ summing to one. $W$ is a $(m \times K)$ matrix
that consists of the mean rewards of the arms for each of the $m$ latent contexts.

**Movie Recommendation Example:** Consider a stylized example of recommending a
Table 1: Example of $\mathbf{W}$ matrix which gives the preferences (rewards) for different genres of movies (arms) based on the latent context

|        | Animation | Action | Family Drama |
|--------|-----------|--------|--------------|
| Adult  | 0         | 0.2    | 0.7          |
| Teenager | 0.2      | 0.8    | 0            |

Table 2: Example of $\mathbf{A}$ matrix which gives the viewership percentage (convex combinations) for different user accounts

| User account | Adult | Teenager |
|--------------|-------|----------|
| User account 1 | 30 %  | 70 %     |
| User account 2 | 0 %   | 100 %    |
| ...          | ...   | ...      |
| User account 100 | 40 %  | 60 %     |

movie to an user account in Netflix, where each user account is shared by various family members in a household. Here, the number of user accounts is $L = 100$ (observed contexts), and the number of movies types is $K = 3$ (arms). Also for simplicity, assume that the chance of liking a particular genre of movie depends only whether the person viewing is an adult or teenager. Thus, the label – adult or teenager – serves as the latent context ($m = 2$), and Table 1 is the $\mathbf{W}$ matrix which encodes the mean rewards (chance of liking the recommendation) for each latent context. Now, each user account is potentially different according to the proportion of viewing times shared by adults and teenagers in that family. This aspect is captured by Table 2 which shows the percentage of viewership by teenagers and adults in each family. This table represents the matrix $\mathbf{A}$, which specifies the relationship between latent contexts and each user account. It is important to note that in this example, Netflix does not directly know either $\mathbf{A}$ or $\mathbf{W}$ – these need to be learned online from data.

**Online Sampled NMF Problem:** Coming back to the general setting, the matrix $\mathbf{U}$ consists of convex combinations of extreme points in the positive orthant. This is identical to the NMF setting. However, unlike the standard NMF problem, we do not get to view either the entire matrix $\mathbf{U}$ or even a noisy version of it. We are instead presented with a context $s \in [L]$ at each time, whose mean rewards corresponds to a row in $\mathbf{U}$. Crucially, this context is randomly chosen from a distribution that we do not control (thus the row of $\mathbf{U}$ that is presented is not under our control). Further, we do not get to observe even this row (of size $K$) – instead, we can choose one coordinate $k$ in this row $s$ to sample, and get a noisy observation of this coordinate (e.g. Bernoulli) with mean $U_{sk}$.

### 1.2 Main Contributions

We propose a latent contextual bandit model where there are $L$ observed contexts and $K$ arms; further the mean rewards of the arms for each observed context can be expressed as a convex combination of the rewards of a small number of ($m$) latent contexts. The main contributions are as follows.

1. **(NMF-Bandit Algorithm)** We propose a latent contextual bandit algorithm that, in an online fashion, multiplexes two tasks. The **first task** refines the current estimate of matrix $\mathbf{A}$ by performing a non-negative matrix factorization (NMF) on the sampled version of a carefully chosen sub-matrix of the mean-reward matrix $\mathbf{U}$. The **second task** uses the current estimate of $\mathbf{A}$ and refines the estimate of $\mathbf{W}$ from sampled versions of several sub-matrices of
In the first task, our sampling projects the rows of $U$ onto lower dimensions such that
the extreme points of $W$ are preserved after the projection. To the best of our knowledge
there has been no prior work on NMF from incomplete data, with theoretical guarantees.
In fact, from existing noisy matrix completion literature, it can be observed that one of
the key conditions to derive spectral norm bounds between the recovered matrix and the
ground truth is that the noise in each entry should be $O(1/K)$ in a $L \times K$ matrix [16].
In the bandit setting where errors occur due to sampling, this would lead to a regret of at
least $O(LK \log t)$. In contrast, our algorithm has much stronger regret guarantees that scale
as $O(L \text{poly}(m, \log K) \log t)$. We show that our algorithm succeeds when the non-negative
matrices $A$ and $W$ satisfy some conditions similar to RIP (Restricted Isometry Property)
and REC (Restricted Eigenvalue Condition) in compressive sensing literature [11]. Further,
we prove a lower bound for this setting which is only $\text{poly}(m, \log K)$ factors away from our
upper bound.

2. (Generative Models for $A$ and $W$) We propose a general family of generative
models for the factors $A$ and $W$ which satisfy the above sufficient conditions for recovery.
These models are extremely flexible, and employ a random + deterministic composition,
where there can be large number of arbitrary bounded deterministic entries (see Section 2.4
for details). The remaining random entries in the matrices are generated from mean-shifted
sub-gaussian distributions (commonly used in the compressive sensing literature [11]).

Finally, we numerically compare our algorithm with contextual versions of UCB-1 and
Thompson Sampling algorithms on synthetic data-sets.

1.3 Related Work

The current work falls at the intersection of learning of low-dimensional structures and
multi-armed bandit problems. We briefly review the areas of contextual bandit literature
and the NMF community that are most relevant to our work.

**Contextual Bandit Problems:** There has been significant progress in contextual
bandits both in the adversarial setting and in the stochastic setting. In the adversarial setting,
an adversary is free to choose the rewards of each arms at each time step. In this setting
the best known regret bounds scale as $O(\sqrt{LKT \log K})$ [6, 29] where $L$ is the number of
contexts and $K$ is the number of arms. In the stochastic regime where there is a constant
gap from the best arm, it can be shown that the regret scales as $O(LK \log t)$ [30]. In [25]
contextual bandits with similarity information in a metric space has been studied both in the
adversarial and the stochastic setting. Contextual bandits with linear payoff functions has
been analyzed in [2, 7] in the adversarial setting while in [1], this has been analyzed in the
stochastic setting. In [10] the authors have expanded this model to the generalized linear
model regime. However, these models require low dimensional features to be known a-priori;
in contrast, our algorithm jointly learns underlying structure in the sampled data.

A related line of work is the online clustering of bandits [12, 21]. The most related in
this line is [12] where each user lies in an (unknown) cluster that can be described by a
low-dimensional vector, with $\ell_2$ separation across clusters. Further, items (arms) have known
features. With this setup, [12] studies joint clustering and regret minimization. A similar
setting is also considered in [31] where arms have known feature vectors, each user lies in one
of a finite number of classes and each user class shares the same user feature vector. In these works, the reward of an arm is a noisy dot-product between the unknown user feature vector and the known arm feature vector. In contrast (and at a high-level), our work differs in that the observed contexts are mixtures of underlying latent structure, and we also do not know the latent structure of the arms (i.e. both observed contexts and arms have latent unknown structure).

**Online Non-negative Matrix factorization (NMF):** The NMF problem has generated a lot of interest in the area of semi-supervised topic modeling as it is useful in generating sparse set of relevant features from non-negative data [13]. However, the NMF problem in the most general form is NP-hard [28]. Arora et. al. have shown that if the matrix is separable and has some robustness properties [3], then NMF is solvable efficiently. Since then, there has been a lot of work in proposing efficient scalable algorithms for NMF, out of which [14, 8, 23] are of particular interest. All these algorithms tackle NMF by solving simple linear programs. There has been some progress in online NMF [9, 15] which aims to update the features efficiently in a streaming sense. However, to the best of our knowledge there has been no work in NMF with bandit feedback with theoretical guarantees. Also related is Kawale et. al. [20] where they propose a Thompson sampling algorithm for online matrix factorization, however they only have theoretical analysis for the rank 1 case.

**Counterfactual analysis:** Recently, contextual bandit algorithms have found use within the framework of causality. In [4], the authors investigate a simple causal graph of three variables $X, Y, Z$, where $X, Y$ are observed, and $Z$ is a latent confounder that causes both, and $X$ causes $Y$. (See [4] and [22] for related definitions.) In the special setting of [4], authors can sample from the counterfactual distribution. This problem is similar to the latent contextual bandit problem we consider in this paper: $W$ depends primarily on the latent variable $Z$, which in turn influences the rewards for all the observed contexts encoded in the matrix $U$. However, [4] does not consider our scaling regime nor provide theoretical guarantees (and has a very different algorithm).

## 2 Problem Statement and Main Results

### 2.1 System Model

**Latent and Observed Contexts:** We consider a stochastic bandit model with $K$ arms and $L$ observed contexts. The set of arms is denoted by $A = \{1, 2, \ldots, K\}$ and the observed contexts belong to the set $S = \{1, 2, \ldots, L\}$. The observed contexts are correlated and depend on a set of latent contexts $\mathcal{Z} \subset S$ where $|\mathcal{Z}| = m << K < L$, with $L = \Omega(K \log K)$. Let $\mathcal{Z} = \{z_1, z_2, \ldots, z_m\}$. A natural interpretation is that, at any time nature chooses a latent context $z \in \mathcal{Z}$, and based on that, a context $s \in S$ is actually observed. We denote the posterior probability of a latent context $z$ given an observed context $s$ as,

$$
P(Z = z_i | S = s) = \alpha_{si}, \forall s \in S \setminus \mathcal{Z}, z_i \in \mathcal{Z} \text{ and } \alpha_{sj} = 1 \{s = z_j\} \forall s, z_j \in \mathcal{Z}
$$

Let $A$ be the matrix with elements $\alpha_{si}$ where $s \in \{1 \cdots L\}$ and $i \in \{1, 2 \cdots m\}$. Please note that the sub-matrix corresponding to the row indices in $\mathcal{Z}$ from an identity matrix $I_{m \times m}$. We
also define the marginal probability of observing a context \( s \in S \) as \( \mathbb{P}(S = s) = \beta_s, \ \forall s \in S \). This specifies the joint distribution of the latent context \( Z \) and the observed context \( S \).

**Bandit Setting:** In this setting the contextual bandit problem can be described as follows: (i) At each time \( t \) the algorithm observes a context \( S_t = s_t \in S \); (ii) After observing the context the algorithm selects an arm \( X_t = x_t \in \mathcal{A} \); and (iii) The algorithm then obtains a Bernoulli reward \( Y_t \) with mean \( U_{x_t,s_t} \). The mean rewards \( U_{x_t,s_t} \) have a latent structure described in the next subsection.

**Rewards:** When an observed context \( s \) is provided, the reward for arm \( k \) depends only on the latent variables. Consider an \( m \times K \) reward matrix \( W \). \( W_{ik} \) specifies the mean reward for arm \( k \) when the latent context is \( z_i \). For all observed contexts \( s \in S \), the mean rewards are given by the matrix \( U \). This is given by:

\[
U_{sk} = \sum_i \mathbb{P}(Z = z_i | S = s) W_{ik} = \sum_i \alpha_{si} W_{ik}.
\]

Therefore, we have \( U = AW \). Since the latent contexts \( Z \) are also a subset of observed contexts, the matrix \( A \) contain a \( I_{m \times m} \) sub-matrix. This is equivalent to the separability condition and is widely used in the NMF literature (see [14]).

**Regret:** The goal is to minimize regret (also known as pseudo-regret [6]) when compared to a genie strategy which knows the matrix \( U \). Let us denote the best arm under a context \( s \in S \) by \( k^*(s) \) and the corresponding reward by \( u^*(s) \). Now, we are at a position to define the regret of an algorithm at time \( T \),

\[
R(T) = \sum_{s \in S} \sum_{\{t \in [T]: S_t = s\}} (u^*(s) - \mathbb{E}[Y_t])
\]

Note that the genie policy always selects the arm \( k^*(s) \) when \( S_t = s \). The class of policies we optimize over are agnostic to the true reward matrix \( U \) and \( Z \), however we assume that \( m \) (the latent dimension) is a known scalar parameter.

### 2.2 Notation

We denote matrices by bold capital letters (e.g. \( U \)) and vectors with bold small letters (e.g. \( x \)). For a \( L \times K \) matrix \( U_{S,:} \) denotes the sub-matrix restricted to the rows in \( S \subset [L] \), while \( U_{:,R} \) denotes the sub-matrix restricted to the columns in \( R \subset [K] \). \( \sigma_m(\mathbf{P}) \) denotes the \( m \)-th smallest singular value of \( \mathbf{P} \). \( \|x\|_p \) denotes the \( \ell_p \)-norm of \( x \). For a matrix \( \|U\|_{\infty,1} \) refers to the maximum \( \ell_1 \)-norm among all the rows while \( \|U\|_2 \) and \( \|U\|_F \) denotes its spectral and Frobenius norms respectively. \( \|U\|_{\infty,\infty} \) denotes the maximum absolute value of an element in the matrix. \( \text{Ber}(p) \) denotes a Bernoulli random variable with mean \( p \).

### 2.3 Sufficient Conditions and Main results

Let us first go over a few definitions before delving into our main results.

**Definition 1.** Consider an \( m \times m' \) matrix \( \mathbf{P} \) with \( m' \geq m \). Define \( \psi_m(\mathbf{P}) = \inf_{a \neq 0: a^T 1 = 0} \frac{\|a^T \mathbf{P}\|_2}{\|a\|_2} \).
Definition 2. Consider an $m \times m'$ matrix $P$ with $m' \geq m$. Define $\psi_m^1(P) = \inf_{a \neq 0: a^T 1 = 0} \|a^T P\|_1 \|a\|_1$.

In what follows we assume $W$ and $A$ satisfy the $\ell_1$-RRSVP and $\ell_2$-RSVP respectively which are randomized variants of the well-known RIP and REC conditions [11]. Note that in Section 2.4 we provide very reasonable generative models that satisfy these conditions with high probability.

Definition 3. (Random Singular Value Property – RSVP) An $L \times m$ matrix ($L \geq m$) $P$ satisfies the $\ell_2$-random singular value property ($\ell_2$-RSVP) with constants $(\epsilon, \rho, m')$ if $\Pr|S| = m' (\sigma_m(P_S) \geq \rho) \geq 1 - \epsilon$ where the probability is taken over sampling a set $S$ of size $m'$ uniformly from $[L]$.

Definition 4. (Random Restricted Singular Value Property – RRSVP) An $m \times K$ matrix ($K \geq m$) $P$ satisfies the $\ell_1$-random restricted singular value property ($\ell_1$-RRSVP) with constants $(\epsilon, \rho, m')$ if $\Pr|S| = m' (\psi_m^1(P_S, S) \geq \rho) \geq 1 - \epsilon$ where the probability is taken over sampling a set $S$ of size $m'$ uniformly from $[K]$.

Now we are at a position to state Theorem 1 which hints at the existence of an algorithm for the latent contextual bandit setting, with regret that scales at a much slower rate than the usual $O(LK \log t)$ guarantees.

Theorem 1. Consider the bandit model with reward matrix $U = AW$. Suppose $A$ is separable [23]. Let $A$ satisfy $\ell_2$-RSVP with constants $(\delta/L, \rho_2, m'_1)$ while $W$ satisfies $\ell_1$-RRSVP with constants $(\delta, \rho_1, m'_2)$. Let $m' = \max(m'_1, m'_2) = \Theta(m \log K)$. Suppose $\beta_s = \Omega(1/L)$ for all $s \in [L]$. We also assume that $L = \Omega(K \log K)$. Then there exists a randomized algorithm whose regret at time $T$ is bounded as,

$$R(T) = O(L\text{poly}(m, \log K) \log t)$$

with probability at least $1 - \delta$, provided the mean of the best arm is separated by a constant gap from the rest of the arms for each context.

We present an algorithm that achieves this performance in Section 3. This theorem is re-stated as Theorem 8 in the appendix which has greater details specific to our algorithm. In the next sections we propose generative models for $A$ and $W$ that satisfy the properties listed above with high probability. We also explain why these generative models are extremely reasonable for our setting.

2.4 Generative Models for $W$ and $A$

The models for $W$ and $A$ are general, and allow for a large number of entries to be arbitrarily deterministic. The rest of the entries are mean shifted, bounded sub-gaussian random variables with some additional mild conditions. Note that this allows sufficient flexibility in our model to represent real world data, for instance each context is likely to have some exclusive arms that have a reward structure represented by the large deterministic part of the $W$ matrix, while other arms may be sufficiently random. Uniform prior on reward that has been used in bandit setting [19] reduces to a special case of this model. Similarly, the
conditions on $A$ allow for a large class of distributions on the $m$-dimensional simplex (with $0 \leq A_{ij} \leq \lambda < 1$) and at the same time allows an identity matrix of dimension $m$ to account for the separability condition [23]. We present the salient features of the model below. We refer to Section 5.2 for a more detailed discussion of the generative models.

1. **Bounded entries**: We ensure that $1 \geq W_{ij} \geq 0$, and $0 \leq A_{ij} \leq \lambda < 1$ a.s.

2. **Random + Deterministic Composition**: Consider a $D \subseteq [K]$ such that $|D| \leq K/(32m)$. $W_{i,D}$ is an arbitrary deterministic matrix (this allows a lot of flexibility – for instance, the maximum entry in each row could be chosen to lie in this deterministic matrix). We allow even more flexibility in $A$ – $A$ has an arbitrary deterministic part $A_{E,:}$, where $E \subseteq [L]$ and the random part $A_{E,:,i}$. Let $|E| \leq \rho L$. $\rho = 1/18$. Row sum of every row of $A_{E,:,i}$ is $1$. The deterministic part of $A$ has an identity of dimension $m$ as a sub-matrix to ensure separability.

3. **Bounded randomness in the random part**: The random entries of both $W$ and $A$ are in “general position”, i.e., they arise from mean shifted bounded sub-gaussian distributions (see Section 5.2, and also [11]) for similar conditions in compressed sensing literature). A few mild conditions are needed on the mean of the entries $E[W_{ij}]$ in the random part to ensure the $\ell_1$-RRSVP condition. The random part of $A$ is a row-normalized version of a random matrix $\hat{A}$. Each entry in $\hat{A}$ needs to satisfy conditions very similar to that of $W$ in order to ensure the $\ell_2$-RSVP condition.

One of our main results is stated as Theorem 2, which implies that if $W$ comes from our generative model then with high probability projecting it onto a small random subset of its columns preserves the $\alpha$-robust simplical property [23] which is a key step in our algorithm.

**Theorem 2.** Let $m' \geq \frac{512}{21e} m \log(eK)$. Let $W$ follow the random model in Section 5.2. $W$ satisfies ($\ell_1$-RRSVP) with constants $(2 \exp(-c_1 \log(eK)), \frac{13}{60} \frac{\sqrt{15m}}{8m}, 2m')$ with probability at least $1 - \exp(-c'_1 \log(eK))$. Here, $c_1, c'_1$ are constants that depend on the sub-gaussian parameter $c(q)$ that depends on the variance in the model for $W$.

In Theorem 3, we follow very similar techniques to prove that small random subsets of rows of $A$ have singular values bounded away from zero with high probability if $A$ is drawn from our generative model.

**Theorem 3.** Let $m' \geq \frac{512}{21e} m \log(eL)$. Let $A$ follow the random model in Section 5.2. $A$ satisfies ($\ell_2$-RSVP) with constants $(2 \exp(-c_2 \log(eL)), \frac{1}{20} \frac{\sqrt{m'}}{m}, 2m')$ with probability at least $1 - \exp(-c'_2 \log(eL))$. Here, $c_2, c'_2$ are constant the depends on the sub-gaussian parameter $c(q)$ that depends on the variance in the model for $A$.

The proof of these theorems are available in the appendix in Section 5.3.

### 2.5 Lower Bound

We prove a problem-specific regret lower bound for a specific class of parameters $(U, W, A)$ which is only a poly$(m, \log K)$ factor away from the upper bound achieved by our algorithm. The lower bound holds for all policies in the class of $\alpha$-consistent policies [24] defined below.
Definition 5. A scheduling policy is said to be $\alpha$-consistent if given any problem instance $U$ we have, $\mathbb{E} \left[ \sum_{t \in [T] : S_t = s} 1 \{ X_t = k \} \right] = O(T(s)^{\alpha})$ for all $k \neq k^*(s)$ and $s \in S$, where $\alpha \in (0, 1)$ and $T(s) = \sum_{t=1}^{T} 1 \{ S_t = s \}$.

Theorem 4. There exists a problem instance $(U, A, W)$ with $\beta_s = \Omega(1/L)$ for all $s \in S$ such that the regret of any $\alpha$-consistent policy is lower-bounded as follows,

$$R(T) \geq (K - 1)mD(U) \left( (1 - \alpha)(\log(T/2m) - \log(L/m)) - \log(4KC) \right)$$

for any $T > \tau$, where $C, \tau$ are universal constants independent of problem parameters and $D(U)$ is a constant that depends on the entries of $U$ and is independent of $L, K$ and $m$.

The proof of this theorem has been deferred to the appendix in Section 5.10 where we specify the class of problem parameters for which we construct this bound.

3 NMF-Bandit Algorithm

In this section we present an $\epsilon$-greedy algorithm (aka the NMF-Bandit algorithm). The NMF-Bandit algorithm takes advantage of the correlation between observed contexts. The algorithm explores with probability $\epsilon_t$; in this case it samples from specific sets of arms (to be specified later). Otherwise w.p. $(1 - \epsilon_t)$ it exploits, i.e., chooses the best arm based on current estimates of rewards to minimize regret. The key steps in the algorithm are as follows.

(a) At each time $t$ and with probability $\epsilon_t$, the algorithm explores, i.e. it randomly performs one of these two steps:

Step 1 – (Sampling for NMF in low dimensions to estimate $A$): Given that it explores, with probability $\alpha$ it samples a random arm from a subset $S \subset [K]$ of arms. $|S| = 2m'$ for $m' = O(m \log L)$. The set $S$ is a randomly chosen at the onset and kept fixed there after.

Step 2 – (Sampling for estimating $W$): Otherwise with probability $(1 - \alpha)$, it samples in a context dependent manner. If the context at the time is $s_t$, the algorithm samples one arm at random from a set of $m$ arms given by $R(s_t)$ (the selection of these sets are outlined below). The context specific sets of arms are designed at the start of the algorithm and held fixed there after.

(b) Otherwise with probability $(1 - \epsilon_t)$ it exploits by performing Step 3 below.

Step 3 – (Choose best arm for current observed context): Suppose the current estimates of $A$ and $W$ from the explore samples are $\hat{A}(t)$ and $\hat{W}(t)$ respectively. Let $\hat{U}(t) = \hat{A}(t)\hat{W}(t)$. The algorithm plays the arm given by $\arg \max_{k \in [K]} \hat{U}(t)_{s_t,k}$, i.e., the best arm for the observed context according to current estimates.

For solving the NMF to obtain $\hat{A}(t)$, we use a robust version of Hottopix [23, 14] as a sub-routine. A detailed pseudo-code of our algorithm has been presented as Algorithm 1 in the appendix. Now, we briefly touch upon the construction of the context specific sets of arms in Step 2 of the explore phase. These sets have been defined in detail in Section 5.1 in the appendix. Let $l = \lfloor K/m \rfloor$. A set $R \subset [L]$ of contexts is sampled at random, such
that $|R| = 2(l+1)m'$ at the onset of the algorithm. We partition $R$ into $l+1$ contiguous subsets $\{S(1), S(2), ..., S(l+1)\}$ of size $2m'$ each. In Step 2 of explore, if $s_t \in S(i)$, then $R(s_t) = \{(i-1)m, (i-1)m+1, \cdots \max(im-1, K)\}$. If $s_t \notin S(i)$ for all $i \in \{l+1\}$, then the algorithm is allowed to pull any arm at random, and these samples are ignored.

### 3.1 Theoretical Insights

A more detailed version of our main result (Theorem 1) has been provided in the appendix (Theorem 8) along with a detailed proof. Theorem 8 exactly specifies the algorithm parameters $\epsilon_t$, $\alpha$ and $m'$ under which we obtain the regret guarantees. Below, we discuss some of the key challenges in the theoretical analysis.

**Noise Guarantees for samples used in NMF:** Matrix completion algorithms that work under the incoherence assumptions require the noise in each element of the matrix to be $O(1/K)$ in order to provide $\infty$-norm guarantees on the recovered matrix [16]. In order to ensure such noise guarantees, we require a very large number of samples in order for estimates to converge. This in turn increases bandit exploration which implies that regret scales as $O(LK \log t)$. To avoid this, we follow a different route. In Step 1 of the explore phase, the NMF-Bandit algorithm only samples from a small subset of arms denoted by $S$. By leveraging the $\ell_1$-RRSVP property of $W$, we can ensure that NMF on these samples (which are basically a noisy version of $U_{:,S}$) gives us a good estimate of $A$ at time $t$; this estimate is denoted by $\hat{A}(t)$. We prove this statement formally in Lemma 6. Given that we sample only from a small subset of arms in the first step of explore, in Lemma 11 we show that the samples concentrate sharply enough.

**Ensuring enough linear equations to recover $W$:** Recall that the reward matrix has the structure $U = AW$. Therefore, an initial approach would be to use the current estimate of $A$ along with samples of the rewards, and directly recover $W$. This however will not work due to lack of concentrations. First, the estimate of $A$ in the early stages will be too noisy to provide sharp estimates about the location of the extreme points aka the latent contexts. Even if we knew the identities of the observed contexts that correspond to “pure” latent contexts (extreme points of the affine space corresponding to the observed contexts), most observed contexts will not correspond to these extreme points – thus, a large number of samples will be wasted, again leading to poor concentrations. Second, if one decides to sample the entries in $U$ at random, the concentration of the entries would be too weak. As before, these weak concentrations will imply $O(LK \log t)$ regret.

Instead, we design the context dependent sets of arms to pull in Step 2 of the explore phase, such that we get enough independent linear equations to recover $W$. The key is to have a small number of arms to sample per observed contexts, but the small number of arms differ across observed contexts. In this case, we show that by leveraging the $\ell_2$-RSVP property of $A$ we can get a good estimate of $W$, denoted by $\hat{W}(t)$ even in the presence of sampling noise. Since we sample from a small subset of arms for each observed context, in Lemma 12 we can ensure that we have sharp concentrations.

**Scheduling the optimal arm during exploit:** The $\infty$-norm bounds on the errors in $\hat{A}(t)$ and $\hat{W}(t)$, imply that $\|\hat{U}(t) - U\|_{\infty,\infty} < \Delta/2$ with probability at least $1 - O(Lm')$ provided $\epsilon_t$ is sufficiently big (see proof of Theorem 8). Here $\Delta = \min_{s \in [L]} (u^*(s) - \max_{k \neq k^*(s)} U_{s,k})$. 

10
This essentially implies that the correct arm is pulled at time $t$ w.h.p if the algorithm decides to exploit.

4 Simulations

We benchmark the performance of our algorithm on synthetic datasets that adhere to the generative models proposed. In Figure 1, we test our algorithm against contextual versions of UCB and Thompson sampling algorithm on a mean reward matrix $U$, where $L = 455$ and $K = 210$ and has non-negative rank 7 in Figure 1a. In Figure 1b we use a reward matrix with $L = 435$ and $K = 105$ and non-negative rank 3. $W$ is generated from an i.i.d distribution uniform over $[0, 1]$. However, after generating the matrix we modify a few entries to ensure that $\Delta$ is 0.1. Note that this falls well within the purview of our generative model for $W$. Each row of $A$ is generated uniformly from the $m$-dimensional simplex. We set $U = AW$. In Figure 1a rewards are generated from an uniform distribution around the means, while in Figure 1b the rewards are drawn from Bernoulli distributions with the mean parameters governed by $U$. It can be observed that our algorithm significantly outperforms both the contextual bandit algorithms; specifically when the number of contexts and arms are large, UCB-1 and Thompson sampling do not collect enough samples to enter the logarithmic regime.

![Figure 1: Comparison of contextual versions of UCB-1 and Thompson sampling with Algorithm 1 (NMF-Bandit) in a latent contextual setting. In (a) Number of observed contexts ($L$) = 455, Number of arms ($K$) = 210 and Number of latent contexts ($m$) = 7. The rewards are Uniform around the means with a support of length 0.4. In (b) Number of observed contexts ($L$) = 435, Number of arms ($K$) = 105 and Number of latent contexts ($m$) = 3. The rewards are Bernoulli with the given means.](image)

5 Appendix – Algorithm and Proof Details

5.1 Algorithmic Details

We present a precise version of the algorithm described in Section 3 as Algorithm 1. For ease of exposition, we introduce the concept of matrix sampling, which is a notational tool to
we pull an arm uniformly at random from
Then we have
vector estimate
the algorithm is allowed to pull any arm at random, and these samples are ignored.
then $R$ set $S$ is sampled at random, such that the subset relevant to the context specific sampling procedure in Step 2 of the subset $R$ is sampled at random, such that $R(s_t)$ is sampled at random, such that $|R| = 2(l + 1)m'$ at the onset of the algorithm. We partition $R$ into $l + 1$ contiguous subsets $\{S(1), S(2), \ldots, S(l + 1)\}$ of size $2m'$ each. The elements of the set $S(j)$ will be denoted as $S(j) = \{s_1(j), s_2(j), \ldots, s_{2m'}(j)\}$. In Step 2 of explore, if $s_t \in S(i)$, then $R(s_t) = \{(i - 1)m, (i - 1)m + 1, \ldots, \max(im - 1, K)\}$. If $s_t \notin S(i)$ for all $i \in [l + 1]$, then the algorithm is allowed to pull any arm at random, and these samples are ignored.

1. $G(0)$: An $K \times 2m'$ random matrix formed as follows: An $2m'$ subset $a_1, a_2 \ldots a_{2m'} \subset [K]$ is chosen randomly uniformly among all $2m'$-subsets of $[K]$ and $G(0)_{a_i,i} = 1$, $\forall 1 \leq i \leq 2m'$ and all other entries are 0.

2. $G(i)$: An $K \times m$ matrix such that,
\[
G(i)_{kj} = \begin{cases} 
1, & \text{if } k = (i - 1)m + j \text{ for } j \in \{1, \ldots, m\} \\
0, & \text{otherwise}
\end{cases}
\]
when $i \in \{1, 2, \ldots, l\}$.

3. $G(l + 1)$: An $K \times r$ matrix defined as follows:
\[
G(l + 1)_{kj} = \begin{cases} 
1, & \text{if } k = (lm + j) \text{ for } j \in \{1, \ldots r\} \\
0, & \text{otherwise}
\end{cases}
\]

5.1.1 Matrix Sampling
Consider the $L \times K$ reward matrix $U$. Consider a ‘sampling matrix’ $G$ with dimensions $K \times p$. Let $\{a_1, a_2, \ldots a_p\} \subset [K]$. In this work, we consider $G$ only of the following form: $G_{a_i,i} = 1$, $\forall 1 \leq i \leq p$ and zero otherwise. Consider the product between a row $s$ of $U$ and $G$, i.e. $U_s \cdot G$. This selects the co-ordinates corresponding to $\{a_1, \ldots a_p\}$ in vector $U_s$. Given a row $s$ (a context $s$) of $U$, i.e. $U_s := \hat{u}[s]$, we describe how to obtain a random Bernoulli vector estimate $\hat{u}[s]$ such that $\mathbb{E}[\hat{u}[s]] = \frac{1}{p}U_s$: by sampling an arm as follows:

- Given that the context is $s$, sample a uniform random variable $\kappa$ with support $\{a_1, \ldots a_p\}$, which represents the arm to be pulled after observing the context.
- Conditioned on $\kappa = k$, pull arm $k$ and observe the reward $Y_k \in \{0, 1\}$.
- The random vector sample is then given by $\hat{u}[s]_k = Y_\kappa e_\kappa$.

Then we have $\mathbb{E}[\hat{u}[s]_k] = \mathbb{E}[\mathbb{E}[Y_k | \kappa = k]] = \frac{1}{p}u[s]_k$. In other words, whenever the context is $s$, we pull an arm uniformly at random from $\{a_1, a_2 \ldots a_p\}$ and the samples are collected in $\hat{u}[s]$.

5.1.2 Arms to be sampled during explore
Before we present the pseudocode, we define the sampling matrices $\{G(0), G(1), \ldots, G(l + 1)\}$. Recall that any subset of arms can be encoded in a sampling matrix. $G(0)$ corresponds to the subset $S$ in Step 1 of explore stated in Section 3. For ease of reference, we restate the sets relevant to the context specific sampling procedure in Step 2 of explore. $G(i)$ corresponds to the subset $R(s_t)$ is $s_t \in S(i)$. Let $l = \lfloor K/m \rfloor$ and $r = K \mod(m)$. A set $R \subset [L]$ of contexts is sampled at random, such that $|R| = 2(l + 1)m'$ at the onset of the algorithm. We partition $R$ into $l + 1$ contiguous subsets $\{S(1), S(2), \ldots, S(l + 1)\}$ of size $2m'$ each. The elements of the set $S(j)$ will be denoted as $S(j) = \{s_1(j), s_2(j), \ldots, s_{2m'}(j)\}$. In Step 2 of explore, if $s_t \in S(i)$, then $R(s_t) = \{(i - 1)m, (i - 1)m + 1, \ldots, \max(im - 1, K)\}$. If $s_t \notin S(i)$ for all $i \in [l + 1]$, then the algorithm is allowed to pull any arm at random, and these samples are ignored.

1. $G(0)$: An $K \times 2m'$ random matrix formed as follows: An $2m'$ subset $a_1, a_2 \ldots a_{2m'} \subset [K]$ is chosen randomly uniformly among all $2m'$-subsets of $[K]$ and $G(0)_{a_i,i} = 1$, $\forall 1 \leq i \leq 2m'$ and all other entries are 0.

2. $G(i)$: An $K \times m$ matrix such that,
\[
G(i)_{kj} = \begin{cases} 
1, & \text{if } k = (i - 1)m + j \text{ for } j \in \{1, \ldots, m\} \\
0, & \text{otherwise}
\end{cases}
\]
when $i \in \{1, 2, \ldots, l\}$.

3. $G(l + 1)$: An $K \times r$ matrix defined as follows:
\[
G(l + 1)_{kj} = \begin{cases} 
1, & \text{if } k = (lm + j) \text{ for } j \in \{1, \ldots r\} \\
0, & \text{otherwise}
\end{cases}
\]
In words, \( G(i) \) for \( i \in [l] \) is the \( K \times m \) matrix which has an identity matrix \( I_{m \times m} \) embedded between rows \((i - 1)m \) and \( im - 1 \), and is zero everywhere else.

### 5.1.3 Representation of the collected Samples

In what follows, let the mean of samples collected through \( G(0) \) till time \( t \) be collected in a \( L \times 2m' \) matrix \( \hat{F}'(t) \) such that \( \mathbb{E} \left[ \hat{F}'(t) \right] = (1/2m')F = (1/2m')UG(0) \) as detailed in Section 5.1.1. Let \( \hat{F}(t) = 2m'\hat{F}'(t) \). Let the samples collected from \( G(i) \) be stored in a \( 2m' \times m \) matrix \( \hat{M}'_i(t) \) such that \( \mathbb{E} \left[ \hat{M}'_i(t) \right] = \frac{1}{m}A_{S(i)}:WG(i) \) for all \( i \in \{1, 2, ..., l + 1\} \). Let \( \hat{M}_i(t) = m\hat{M}'_i(t) \) be the scaled version.

### 5.1.4 Pseudocode

**Algorithm 1** NMF-Bandit - An \( \epsilon \)-greedy algorithm for Latent Contextual Bandits

1. At time \( t \),
2. Observe context \( S_t = s_t \)
3. Let \( E(t) \sim \text{Ber}(\epsilon_t) \)
4. if \( E(t) = 1 \) then
5. \hspace{1em} **Explore**: Let \( H_t \sim \left\{ \begin{array}{ll}
\text{Ber}\left(\frac{2m'}{r+2m'}\right), & \text{if } s_t \in S(l + 1) \\
\text{Ber}\left(\frac{2m'}{m+2m'}\right), & \text{otherwise}
\end{array} \right. \)
6. \hspace{1em} If \( H_t = 1 \) sample an arm according to the matrix sampling technique applied to matrix \( G(0) \) and update \( F(t) \).
7. \hspace{1em} If \( H_t = 0 \) sample an arm according to the matrix sampling technique applied to matrix \( G(i) \) if \( s_t \in S(i) \) for \( i \in \{1, 2, ..., l + 1\} \) and update \( \hat{M}_i(t) \). If \( s_t \) is not in any of these sets then choose an arm at random.
8. else
9. \hspace{1em} **Exploit**:
10. \hspace{2em} Let us compute,
11. \hspace{2em} \( \hat{W}(t) = \text{Hottopix}(F(t), m, 2m'\gamma(t)) \).
12. \hspace{2em} \( \hat{A}(t) = \arg\min_{Z \geq 0, \text{rowsum}(Z) = 1} \left\| F(t) - Z\hat{W}(t) \right\|_{\infty, 1} \).
13. \hspace{2em} Let \( \hat{W}(t) \in \mathbb{R}^{m \times K} \) be such that,
14. \hspace{2em} \( \hat{W}(t)_{:, (i-1)m:im-1} = \arg\min_{X_{m \times m}} \left\| \hat{A}(t)_{S(i)}, X - \hat{M}_i(t) \right\|_2, \forall i \in \{1, 2, ..., l\} \)
15. \hspace{2em} \( \hat{W}(t)_{:, im:K} = \arg\min_{X_{m \times r+1}} \left\| \hat{A}(t)_{S(l)}, X - \hat{M}_{l+1}(t) \right\|_2 \).
16. end if

For the sake of completeness we include the robust version of the Hottopix algorithm [14] which is used as a sub-program in Algorithm 1. The following LP is fundamental to the
Hottopix algorithm,

\[
\min_{C \in \mathbb{R}^{f \times n}} \mathbf{p}^T \text{diag}(C) \\
\text{s.t. } \| \tilde{X} - C \tilde{X} \|_{\infty,1} \leq 2\epsilon \\
\text{and } C_{ii} \leq 1, C_{ji} \leq C_{ii} \forall i, j \in [L]
\]  

where \( \mathbf{p} \) is a vector with distinct positive values.

\section*{Algorithm 2 Hottopix(\( \tilde{X}, m, \epsilon \))}

1: \textbf{Input :} \( \tilde{X} \) such that \( \tilde{X} = AW + N \), where \( A \in [0, 1]^{L \times m} \) and \( \| A_{i,} \|_1 = 1 \) for all \( i \in [L] \), 
\( W \in \mathbb{R}^{m \times 2m'} \) and \( \| N \|_{\infty,1} \leq \epsilon \).
2: \textbf{Output :} \( \hat{W} \) such that \( \hat{W} \sim W \).
3: Compute an optimal solution \( C^* \) to (4).
4: Let \( K \) denote the set of indices \( i \) for which \( C^*_{ii} \geq \frac{1}{2} \).
5: Set \( \hat{W} = \tilde{X}_K, : \).

\section{5.2 Detailed Generative Models for W and A}

The model for \( W \) and \( A \) are both very similar with deterministic and random parts, in fact the only difference comes from the fact that the rows of \( A \) are in the interior of the \( m \)-dimensional simplex which is crucial to the latent structure. We assume that \( A \) and \( W \) satisfy the following properties:

1. \textbf{Bounded entries:} We ensure that \( 1 \geq W_{ij} \geq 0 \), and \( 0 \leq A_{ij} \leq \lambda < 1 \) a.s.

2. \textbf{Random+Deterministic Composition:} Consider a \( D \subseteq [K] \) such that \( |D| \leq K/(32m) \). \( W_{i,D} \) is an arbitrary deterministic matrix. The maximum entry in every row of \( W \) is assumed to be contained in the deterministic part. We allow even more flexibility in \( A \), such that \( A \) has an arbitrary deterministic part \( A_{E,:} \) where \( E \subseteq [L] \) and the random part \( A_{E_c,:} \). Let \( |E| \leq \rho L. \rho = 1/18. \) Row sum of every row of \( A_{E_c,:} \) is 1.

3. \textbf{Bounded randomness in the random part:} For \( j \in D^c \), \( W_{ij} \) is an independent random variable satisfying \( \mathbb{E}[W_{ij}] = m_{ij} \), \( \text{Var}[W_{ij}] = q \) and \( 0 \leq W_{ij} \leq 1 \) a.s. Further, we assume that the matrix \( M_{i,D^c} \) consisting of the means \( m_{ij}, j \in D^c \) is such that \( \sum_{i\neq1} |m_{ij} - m_{ij}|^2 \leq \frac{1}{25}, \forall j \in D^c, i_1 \neq i_2. \) Essentially, this means that:

\[
W_{i,D^c} = 1 \ast \mathbf{m}_{D^c}^T + R_{i,D^c} + \tilde{W}_{i,D^c} \quad (5)
\]

where each \( i,j \)th entry of \( \tilde{W}_{i,D^c} \) is an independent mean zero sub-gaussian entry with variance \( q \), and bounded support and sub-gaussian parameter \( c(q) \). \( \mathbf{m}_{D^c} \in \mathbb{R}^{D^c} \) is an arbitrary deterministic vector where \( \mathbf{m}_{D^c} \) is set to \( M_{1,D^c} \). \( R_{i,D^c} \) is a deterministic perturbation matrix satisfying \( \| R_{i,j} \|_2 \leq \frac{1}{5}, \forall j \in D^c. \) We follow a similar model for
A with minor changes owing to the fact that its rows lie on the simplex. \( A_{E^c} \) is a random matrix \( \tilde{A} \) normalized row-wise. We first describe the random model on \(|E^c| \times m\) matrix \( \tilde{A} \). \( \tilde{A}_{ij}, i \in |E^c| \) is an independent random variable satisfying \( \mathbb{E}[\tilde{A}_{ij}] = n_{ij} \), \( \text{Var}[\tilde{A}_{ij}] = q \) and \( 1/m \leq \tilde{A}_{ij} \leq \gamma < 1 \) a.s., which is to ensure the bounded entries. We denote the matrix of means by \( N \) consisting of the means \( n_{ij} \). The \( \ell_2 \) norm of every row of \( N \) is at most \( 1/5 \). Like in the case for model of \( W \), \( \tilde{A} = N + \hat{A} \) (6)

where \( \hat{A} \) is a matrix with independent mean zero sub-gaussian entries each with variance \( q \), and bounded support and sub-gaussian parameter \( c(q) \). In order to ensure separability \([23]\) we assume that there is a subset \( M \subseteq E : |M| = m \) such that \( A_M = I_{m \times m} \).

### 5.3 Projection onto a Low Dimensional Space

In this section, we will prove some properties of the matrix \( F = UG(0) = AWG(0) \) where \( G(0) \) is a \( K \times 2m' \) as defined in Section 5.1.1. From the definition in Section 2.1, \( A \) contains a \( I_{m \times m} \) sub-matrix corresponding to the rows in \( Z \). Further, the row sum of every row of \( A \) is 1. This means that the rows of \( U \) consists of points in the convex hull of extreme points, i.e. the rows of \( W \), together with the extreme points themselves.

The extreme points in \( W \) are mapped to extreme points in \( WG(0) \). We also show that the new set of extreme points \( WG(0) \) also satisfy what is called the simplical property when \( W \) satisfies the assumptions in Section 5.2.

When the entries in \( W \) are random and independent bounded random variables as in Section 2.4, we show that \( \ell_1 \) distance of any non-zero vector \( a \) such that \( a^T 1 = 0 \) is preserved under the map \( a^T WG(0) \) with high probability over \( W \) for any fixed \( G(0) \). We need some results relating to sub-gaussianity of the matrix \( W \) which we deal with in the next subsection.

### 5.4 Properties of Sub-Gaussian Matrices

**Definition 6.** [11] A random variable \( X \) is sub-gaussian with parameter \( c > 0 \) if \( \mathbb{E}[\exp(tX)] \leq \exp(-c^2t^2), \forall t \in \mathbb{R} \).

**Definition 7.** [11] A random vector \( Y \in \mathbb{R}^n \) is isotropic if \( \mathbb{E}[(Y^Tx)^2] = \mathbb{E}[x^T x], \forall x \in \mathbb{R}^n \). It is sub-gaussian with parameter \( c \) if the scalar random variable \( Y^Tx \) is sub-gaussian with parameter \( c \) for all \( x \in \mathbb{R}^n : \|x\|_2 = 1 \), i.e. \( \mathbb{E}[\exp(t(Y^Tx))] \leq \exp(-ct^2), \forall t \in \mathbb{R}, \forall \|x\|_2 = 1 \).

**Lemma 1.** [11],[26] Consider a random variable \( X \) such that \( \mathbb{E}[X] = 0, \mathbb{E}[X^2] = 1 \), \( |X| \leq b \) a.s for some constant \( b > 0 \). Then, \( X \) is sub-gaussian with parameter \( \frac{b^2}{2} \). Consider a random vector \( Y \in \mathbb{R}^n \) where each entry is drawn i.i.d from a mean zero, unit variance and a sub-gaussian distribution with parameter \( c \). Then \( Y \) is a sub-gaussian isotropic vector with the same sub-gaussian parameter \( c \).

**Remark:** The first part is from Theorem 9.9 in [26] while the second part is from Lemma 9.7 from [11].
Lemma 2. [11] Let $P$ and $Q$ be two matrices of the same dimensions. Let $\sigma_{\text{min}}$ and $\sigma_{\text{max}}$ be the largest and smallest singular values of a matrix respectively. Then,

$$|\sigma_{\text{min}}(P) - \sigma_{\text{min}}(Q)| \leq \sigma_{\text{max}}(P - Q)$$  \hspace{1cm} (7)

Let $P \in \mathbb{R}^{p \times q}$ where $p \geq q$. Then,

$$\sigma_{\text{max}}(P^T P - I_{n \times n}) \leq \delta \Rightarrow \sigma_{\text{min}}(P) \geq \sqrt{(1 - \delta)}$$  \hspace{1cm} (8)

Lemma 3. [11] Consider an $m \times s$ matrix $P$ with every row being a random independent sub-gaussian isotropic vector with sub-gaussian parameter $c$. Let $m > s$, then:

$$\Pr \left( \sigma_{\text{max}} \left( \frac{1}{m} P^T P - I_{s \times s} \right) \geq \delta \right) \leq 2 \exp \left( -\frac{3 \tilde{c}^2 \delta^2 m + 7 s}{2} \right)$$  \hspace{1cm} (9)

Further,

$$\Pr \left( \sigma_s(P) \leq \sqrt{m} \sqrt{(1 - \delta)} \right) \leq \Pr \left( \sigma_{\text{max}} \left( \frac{1}{m} P^T P - I_{s \times s} \right) \geq \delta \right) \leq 2 \exp \left( -\frac{3 \tilde{c}^2 \delta^2 m + 7 s}{2} \right)$$  \hspace{1cm} (10)

Here, $\tilde{c}$ is a constant that depends only on the sub-gaussian parameter $c$.

**Remark:** The first result follows from equation (9.15) in [11] and also from combining Lemma 9.8 and Lemma 9.9 in [11]. The second follows from applying Lemma 2

**Definition 8 ([23]).** Let us consider a matrix $M$ which is $p \times q$ where $p \leq q$. Let $m_i \in \mathbb{R}^{1 \times p}$ be the $i$-th row of the matrix $M$. The matrix $M$ is $\alpha$-simplical if $\min_{i \in \{1 \ldots p\} \setminus \{m_i\}} \|m_i - x\|_1 \geq \alpha$. In other words, every row is at least $\alpha$ far away in $\ell_1$ distance from the convex hull of other points.

### 5.5 Results regarding sub-matrices of $W$

The following results hold for $WG(0)$ since $WG(0) = W_{:,S}$ when $S = \{a_1 \ldots a_m\}$ is the set of column indices associated with $G(0)$ as in Section 5.1.

**Theorem 5.** Let $W$ follow the random generative model in Section 2.4. Let $S \subseteq D^c$. Let $|S| = m' \geq \frac{512}{216} m \log(eK)$,

$$\psi_m(W_{:,S}) \geq \left( \frac{11}{20} \right) \sqrt{m'}$$  \hspace{1cm} (11)

with probability at least $1 - \frac{2}{K^2 m'^2}$ over the randomness in $W$. Here, $\tilde{c}$ is a constant that depends on the sub-gaussian parameter $c(q)$ of the distributions in the generative model in Section 2.4.
\textbf{Theorem 6.} Consider a matrix $W$ in Section 2.4, $W = \tilde{W}_i:S + 1m_s^T + R_s$. Here, $\tilde{W}_i:S$ has sub-gaussian entries with parameter $c(q)$, since by Lemma 1, all bounded random variables on support $[-1, 1]$ with zero mean are sub-gaussian and their sub-gaussian parameter depends on the variance. Let $m_s$ refer to the vector restricted to co-ordinate in $S$. Applying Lemma 3 to the sub-gaussian matrix $(m' \times m) \tilde{W}_i:s$ with $m' \geq \frac{512}{21\epsilon} m \log(eK)$ and setting $\delta = 7/16$, we have:

$$\Pr \left( \sigma_m(\tilde{W}_i:s)^T \leq \frac{3}{4}\sqrt{m'} \right) \leq 2 \exp(-\frac{7}{2}m \log(K)) \leq 2K^{-7m/2}. \quad (12)$$

Now, applying Lemma 2, we have:

$$|\sigma_m(R_i:s + \tilde{W}_i:s) - \sigma_m(\tilde{W}_i:s)| \leq \sigma_{\max}(R_i:s) \leq \|R_i:s\|_F \leq \frac{1}{5}\sqrt{m'} \quad (13)$$

Combining the above two equations, we have:

$$\Pr \left( \sigma_m(\tilde{W}_i:s + R_i:s) \leq \left( \frac{3}{4} - \frac{1}{5} \right)\sqrt{m'} \right) \leq 2 \exp(-\frac{7}{2}m \log(K)) \leq 2K^{-7m/2}. \quad (14)$$

For any fixed set of size $S = m'$, We have the following chain:

$$\inf_{a \neq 0: a^T 1 = 0} \frac{\|a^T W_i:s\|_2}{\|a\|_2} = \inf_{a \neq 0: a^T 1 = 0} \frac{\|a^T(1m_s^T + \tilde{W}_i:s + R_i:s)\|_2}{\|a\|_2}$$

$$= \inf_{(a^T 1 = 0)} \frac{\|a^T(\tilde{W}_i:s + R_i:s)\|_2}{\|a\|_2} \geq \sigma_m(R_i:s + \tilde{W}_i:s) \quad (15)$$

\textbf{Theorem 6.} Consider a matrix $W$ with the generative model in Section 2.4. Let $m' \geq \frac{512}{21\epsilon} m \log(eK)$. For any fixed set $S$ of size $2m'$ such that $S_1 = S \cap D, |S_1| \leq \frac{2m'}{16m}$ we have:

$$\psi_m^1(W_i:s) = \inf_{a \neq 0: a^T 1 = 0} \frac{\|a^T W_i:s\|_1}{\|a\|_1} \geq \left( \frac{13}{60} \right) \sqrt{\frac{15m'}{8m}} \quad (16)$$

with probability at least $1 - 2K^{-7m/2}$ over the randomness in $W$. Further, rows of $W_i:s$ is $\psi_m^1(W_i:s)$-simplicial.

\textbf{Proof.} Let $S_2 = S \cap D^c$. Here, $|S_2| \geq 2m'(1 - \frac{1}{16m}) \geq \frac{15m'}{8} \geq \frac{512}{21\epsilon} m \log(eK)$. The first result follows from the following chain:

$$\|a^T W_i:s\|_1 \geq \|a^T[W_i:s_1 W_i:s_2]\|_2 \geq (a) \|a^T W_i:s_2\|_2 - \|a^T W_i:s_1\|_2$$

$$\geq \|a\|_2 \psi_m(W_i:s_2) - \|a\|_2 \sqrt{\frac{m \cdot 2m'}{16m}} \geq \|a\|_1 \sqrt{\frac{15m'}{8m}} \left( \frac{3}{4} - \frac{1}{5} - \frac{1}{\sqrt{15}} \right)$$

$$\geq \left( \frac{3}{4} - \frac{8}{15} \right) \sqrt{\frac{15m'}{8m}} \text{ w.p. } 1 - \frac{2}{K^{7m/2}} \quad (17)$$
Justifications of the above chain are: (a)- Triangle inequality for the norm $\|\cdot\|_2$. (b)- Definition of $\psi_m(\cdot)$ and $\|a^T W_{S_i}\|_2 \leq \|a^T\|_2 \|W_{S_i}\|_F \leq \sqrt{m} \|S_i\|_1 \|a^T\|_2$. (c)- $\|\cdot\|_2 \geq \frac{\|\cdot\|_1}{\sqrt{m}}$ and applying Theorem 5 because $S_2 \subseteq D_c$ and $|S_2| \geq \frac{512}{216} m \log(eK)$.

For the second part, let us denote $r^{-i} \in \mathbb{R}^{1 \times m}$ to be a vector satisfying $\sum_{k \neq i} r^{-i}_k = -1$, $r^{-i}_k \leq 0 \; \forall k \neq i$ and $r^{-i}_i = 1$. It is easy to easy that:

$$\|r^{-i}\|_1 \geq 1. \tag{18}$$

From the definition for an $\alpha$-simplical matrix (Definition 8), it is enough to show that for any $r^{-i}$, $\|r^{-i} W_{\cdot,S}\|_1 \geq \psi^1_m(W_{\cdot,S})$. We prove this as follows:

$$\|r^{-i} W_{\cdot,S}\|_1 \geq \psi^1_m(W_{\cdot,S}) \tag{19}$$

5.5.1 Choosing a good $S$ for $G(0)$

**Lemma 4.** Let $D$ be the set as defined in Section 2.4. Let a random $2m'$-subset $S$ be chosen out of $[K]$ where $m' = \frac{512}{216} m \log(eK)$. Then, $Pr(|S \cap D| \leq \frac{2m'}{10m}) \leq \exp(-c_1 \log(eK))$ for constant $c_1 > 0$ that depends on $\mathring{c}$.

**Proof.** Let $X_1, \ldots X_{2m'}$ be set of indicator functions such that $X_i = 1$ if the $i$-th element in the random subset $S$ chosen uniformly without replacement belongs to $D$ and it is 0 otherwise. Let $Y_1, Y_2 \ldots Y_{2m'}$ be the set of indicator functions such that $Y_i = 1$ (and 0 otherwise) if the $i$-th element in the random multi-set $S$ belongs to $D$ where the multiset elements are chosen independently and uniformly with replacement. It is clear that $E[X_i] = E[Y_i] = \frac{|D|}{K} \geq \frac{1}{32m}$. The moment generating function of the sum of $X_i$’s is dominated by the moment generating function of the sum of $Y_i$’s. Therefore, all concentration inequalities, based on moment generating functions, for variables drawn with replacement holds for variables drawn without replacement [17]. In particular, the following inequality derived from moment generating functions holds [18] for any $\delta > 0$:

$$Pr \left( \sum X_i \geq (1 + \delta)2m \mu \right) \leq Pr \left( \sum X_i \geq (1 + \delta)2m \mu \right) \leq \exp(\delta 2m \mu (1 + \delta)^{(1+\delta)2m \mu} . \tag{20}$$

Let us take $\delta = 1$. Therefore, $Pr \left( |S \cap D| \geq \frac{2m'}{10m} \right) \leq \left( \frac{4}{e} \right)^{-\frac{32 m \log(eK)}{\sqrt{m} \log(4/e)}} \leq \frac{1}{(eK) \sqrt{m} \log(4/e)}$. \hfill \Box

**Proof of Theorem 2.** From Theorem 6 and Lemma 4 we have,

$$\mathbb{E}_W \left[ Pr \left( \psi^1_m(W_{\cdot,S}) < \frac{13}{60} \frac{\sqrt{15m}}{\sqrt{8m}} \right) \right] \leq \exp(-c_1 \log(eK)) + 2K^{-7m/2} \leq 2 \exp(-c_1 \log(eK))$$

18
Now by Markov’s inequality this implies that,

\[
P_W \left[ \left( \mathbb{P}_S \left( \psi_m^1(W, S) < \left( \frac{13}{60} \right) \frac{\sqrt{15m'}}{\sqrt{8m}} \right) \geq 2 \exp(-\frac{c_1}{2} \log(eK)) \right) \right] 
\leq \frac{\exp(-c_1 \log(eK))}{\exp(-\frac{c_1}{2} \log(eK))}
\leq \exp(-\frac{c_1}{2} \log(eK))
\]

This implies the following chain:

\[
P_W \left[ \left( \mathbb{P}_S \left( \psi_m^1(W, S) > \left( \frac{13}{60} \right) \frac{\sqrt{15m'}}{\sqrt{8m}} \right) \leq 1 - 2 \exp(-\frac{c_1}{2} \log(eK)) \right) \right] 
\leq \exp(-\frac{c_1}{2} \log(eK))
\]

\[
\Rightarrow P_W \left[ \left( \mathbb{P}_S \left( \psi_m^1(W, S) > \left( \frac{13}{60} \right) \frac{\sqrt{15m'}}{\sqrt{8m}} \right) \geq 1 - 2 \exp(-\frac{c_1}{2} \log(eK)) \right) \right] 
\geq 1 - \exp(-\frac{c_1}{2} \log(eK))
\]

This proves that with probability at least \(1 - \exp(-\frac{c_1}{2} \log(eK))\) the \(\ell_1\)-RRSVP condition is satisfied with the said parameters.

\[\square\]

### 5.6 Results regarding sub-matrices of \(A\)

We assume that \(A\) satisfies the random generative model in 2.4. We prove some results regarding the minimum singular values of sub-matrices corresponding to columns in set \(S\) which is a mix of random and the deterministic columns. The proofs follow closely those of \(W\) in the previous section.

**Theorem 7.** Let \(A\) follow the random generative model in Section 2.4. Let \(m' \geq \frac{512}{21\epsilon}m \log(eL)\). Fix any set \(S\) of size \(2m'\) such that \(S_1 = S \cap E, |S_1| \leq \frac{2m'}{9}\). Let \(S_2 = S \setminus S_1\). Then, we have:

\[
\sigma_m(A_{S_2}) \geq \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \text{ w.p } 1 - \frac{2}{L^{7m/2}}. 
\]

**Proof.** Let \(\tilde{S}_2\) be the set of rows in the random matrix \(\tilde{A}\) that corresponds to the rows \(S_2\) in \(A\). Here, \(\tilde{A}_{\tilde{S}_2,:}\) has sub-gaussian entries with sub-gaussian parameter \(c(q)\), since by Lemma 1, all bounded random variables on support \([-1,1]\) with zero mean are sub-gaussian and their sub-gaussian parameter depends on the variance.

Therefore, applying Lemma 3 to the sub-gaussian matrix \((|\tilde{S}_2| \times m) \tilde{A}_{\tilde{S}_2,:}\) with \(|\tilde{S}_2| \geq m' \geq \frac{512}{21\epsilon}m \log(eL)\) and setting \(\delta = 7/16\), we have:

\[
\Pr \left( \sigma_m(\tilde{A}_{\tilde{S}_2,:}) \leq \frac{3}{4} \sqrt{m'} \right) \leq 2 \exp \left( -\frac{7m \log(L)}{2} \right) \leq 2L^{-7m/2}. 
\]
Now, consider the following matrix: \( \left[ \frac{1}{m} \left( N\bar{s}_2; + \hat{A}\bar{s}_2; \right) A_{S_1;} \right] \). First, note that according to the model in Section 2.4, rows of \( A_{S_1;} \) sum to 1. Therefore, we have the following chain for any non-zero vector \( a \in \mathbb{R}^{1 \times m} \):

\[
\| [N\bar{s}_2; + \hat{A}\bar{s}_2; A_{S_1;}]a \|_2 \geq \| N\bar{s}_2; + \frac{1}{m} \hat{A}\bar{s}_2; \|_2 a - \| A_{S_1;}a \|_2 \\
\geq \| (N\bar{s}_2; + \hat{A}\bar{s}_2;) a \|_2 - \sqrt{\sum_{i \in S_1} \| A_i \|_2^2} \|a\|_2^2 \\
\geq \| (N\bar{s}_2; + \hat{A}\bar{s}_2;) \|_2 a - \sqrt{2\rho m} \|a\|_2 \\
\geq \| \hat{A}\bar{s}_2; \|_2 \|a\|_2 - \| N\bar{s}_2; \|_2 \|a\|_2 - \sqrt{2\rho m} \|a\|_2 \\
\geq \sigma_m (\hat{A}\bar{s}_2;) - 2m (1 - \frac{1}{9}) \sqrt{\frac{2}{1 - \frac{1}{9}}} \|a\|_2 \\
\geq \left( \frac{3}{4} \frac{1}{5} - \frac{1}{2} \right) \sqrt{\frac{3}{4} \frac{1}{5}} \|a\|_2 \text{ w.p. } 1 - 2L^{-7m/2}. \tag{27}
\]

Now, we normalize the every row of \( [N\bar{s}_2; + \hat{A}\bar{s}_2; A_{S_1;}] \) to get \( [A_{S_2;}A_{S_1;}] = A_S P \) where \( P \) is a permutation matrix. Now, every entry gets scaled by at least \( 1/m \) since rows sum is at most \( m \). Therefore, the minimum singular value scales by at least \( 1/m \). Therefore,

\[
\sigma_m (A_S) = \sigma_m (A_S P) \geq \frac{\sqrt{m'}}{m} \left( \frac{3}{4} \frac{1}{5} - \frac{1}{2} \right) \text{ w.p. } 1 - 2L^{-7m/2}. \tag{28}
\]

\[\square\]

5.6.1 Choosing a good \( S(i) \) for a \( G(i) \)

**Lemma 5.** Let \( E \) be the set as defined in Section 2.4. Let a random \( 2m'\)-subset \( S \) be chosen out of \( [L] \) where \( m' = \frac{\delta}{2} m \log(eL) \). Then, \( \Pr \left( |S \cap E| \leq \frac{2m'}{9} \right) \leq \exp(-c_2 m \log(eL)) \) for constant \( c_2 > 0 \) that depends on \( \bar{c} \).

**Proof.** The proof is identical to the proof of Lemma 4. We just choose \( \mu = \frac{1}{18} \) and \( \delta = 1 \). Therefore we have:

\[
\Pr \left( |S \cap E| \geq \frac{2m'}{9} \right) \leq \frac{1}{(eL)^{\frac{512 \log(1/\bar{c})}{180m}}}. \tag{29}
\]

\[\square\]
Proof of Theorem 3. From Theorem 7 and Lemma 5 we have,

\[ \mathbb{E}_A \left[ \mathbb{P}_S \left( \sigma_m(A_{S,:}) < \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \right) \right] \leq 3 \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \] (30)

Now by Markov’s inequality this implies that,

\[ \mathbb{P}_A \left[ \mathbb{P}_S \left( \sigma_m(A_{S,:}) < \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \right) \geq \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \right] \leq 3 \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \]

This implies the following chain:

\[ \mathbb{P}_A \left[ \mathbb{P}_S \left( \sigma_m(A_{S,:}) < \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \right) \geq 1 - \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \right] \leq 3 \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \] (31)

\[ \Rightarrow \mathbb{P}_A \left[ \left( \mathbb{P}_S \left( \sigma_m(A_{S,:}) < \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \right) \right) \geq 1 - \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \right] \leq 3 \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \] (32)

\[ \Rightarrow \mathbb{P}_A \left[ \left( \mathbb{P}_S \left( \sigma_m(A_{S,:}) > \frac{\sqrt{m'}}{m} \left( \frac{1}{20} \right) \right) \right) \right] \geq 1 - 3 \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \] (33)

This proves that with probability at least \( 1 - \exp \left( -\frac{c'_2 m \log(eL)}{2} \right) \) the \( \ell_2 \)-RSVP condition is satisfied with the said parameters.

5.7 Noisy NMF in Low dimensions

In this section we enhance the guarantees of the robust Hottopix algorithm from [14] provided \( W \) satisfies \( \ell_1 \)-RRSVP and the subset \( S \) chosen by Algorithm 1 is good as in Section 4.

Lemma 6. Suppose \( W \) satisfies \( \ell_1 \)-RRSVP with parameter \((\delta, \rho_1, 2m')\) and the subset \( S \) of its columns \(|S| = 2m'\) satisfies \( \psi^1_m(W_{S,:}) \geq \rho_1 \). Consider a matrix \( \tilde{X} = AW_{:,S} + N \) such that \( \|N\|_{\infty,1} \leq \epsilon \) and \( A \) is separable [23]. Under these assumptions Hottopix(\( \tilde{X}, m, \epsilon \)) returns \( \hat{W} \) such that,

\[ \|\hat{W} - W_{:,S}\|_{\infty,1} \leq \epsilon \] (35)

if \( \epsilon < \frac{\rho_1 (1 - \lambda)}{15} \). Suppose \( \hat{A} = \arg\min_{Z \geq 0, \text{rowsum}(Z) = 1} \|\tilde{X} - Z\hat{W}\|_{\infty,1} \). Then we have,

\[ \|\hat{A} - A\|_{\infty,1} \leq \frac{4\epsilon}{\rho_1 - \epsilon} \] (36)
Proof. Let \( W' = W_{i,S} \) and \( X = AW_{i,S} \). The bound in (6) is immediate from Theorem 2 in [14] as \( W' \) is \( \rho_1 \)-robust simplicial by Theorem 6. We first note that,

\[
\|X - \hat{A}W\|_{\infty,1} \leq \|X - \hat{A}W'\|_{\infty,1} + \|X - \hat{A}W'\|_{\infty,1} + \|AW' - W\|_{\infty,1}
\]

\[
\leq \|A\|_{\infty,1} \|W' - W\|_{\infty,1} + \epsilon \leq 2\epsilon
\]

The first inequality follows from the triangle inequality while the last one holds because \( \|A\|_{\infty,1} = 1 \). Thus, the LP to recover \( \hat{A} \) will always output \( \hat{A} \) with,

\[
\|X - \hat{A}W\|_{\infty,1} = \|AW' - \hat{A}W\|_{\infty,1} \leq 3\epsilon.
\]

(37)

We can apply triangle inequality to get,

\[
\left\| (A - \hat{A}) W' \right\|_{\infty,1} \leq \left\| AW' - \hat{A}W' \right\|_{\infty,1} + \hat{A} \left\| W' - \hat{W} \right\|_{\infty,1}
\]

\[
\leq 3\epsilon + \|A\|_{\infty,1} \left\| W' - \hat{W} \right\|_{\infty,1} + \epsilon \leq 3\epsilon + (1 + \|A - \hat{A}\|_{\infty,1}) \epsilon
\]

(38)

In order to get the desired result we need to lower bound the L.H.S in (38). Note that rowsum \((A - \hat{A}) = 0\). Therefore we have,

\[
\left\| (A - \hat{A}) W' \right\|_{\infty,1} \geq \|A - \hat{A}\|_{\infty,1} \rho_1
\]

(39)

by definition. Combining (39) and (38) we get the required bound.

\[\square\]

5.8 Noisy Recovery of Extreme Points

In this section we assume that \( A \) satisfies the \( \ell_2 \)-RSVP property with parameter \((\delta/L, \rho_2, 2m')\).

Lemma 7. If \( A \) satisfies the \( \ell_2 \)-RSVP property with parameter \((\delta/L, \rho_2, 2m')\) then the sets \( \{S(1), \ldots, S(l + 1)\} \) satisfy,

\[
\sigma_m(A_{S(i),:}) \geq \rho_2, \text{ for all } i \in [l + 1]
\]

with probability atleast \(1 - \delta\) over the randomness in choosing the subsets.

Proof. The proof of this lemma is just an union bound over all the events \( \{\sigma_m(A_{S(i),:}) < \rho_2\} \). Note that by virtue of \( \ell_2 \)-RSVP each of these events is true with probability atmost \(\delta/L\). \[\square\]

If the conditions of the above lemma are satisfied we will call the corresponding sets good. Recall the definition of \( \hat{M}_t(t) \). We will show that if \( \hat{A}(t) \) is close to \( A \) and the matrices \( \hat{M}_t(t) \) are sufficiently close to their means, then we recover \( W \) upto the same accuracy. Let us define \( M_t = \mathbb{E} \left[ \hat{M}_t(t) \right] \).
Lemma 8. Suppose $A$ satisfies the $\ell_2$-RSVP property and \{S(1), S(2), \cdots S(l+1)\} are good in the sense of Lemma 7. Given that $\|\hat{A}(t) - A\|_{\infty,1} \leq \epsilon_1$ and $\|\hat{M}_i(t) - M_i\|_{\infty,\infty} \leq \epsilon_2$ for all $i \in [l+1]$, $\hat{W}(t)$ recovered by Algorithm 1 satisfies,

$$\|\hat{W}(t) - W\|_{\infty,\infty} \leq \frac{m(2\epsilon_1 + 3\epsilon_2)}{\rho_2}$$  \hspace{1cm} (40)

if $\epsilon_1, \epsilon_2 \leq \frac{\rho_2}{m}$.

Proof. Let $\hat{W}(t)_{(i-1)m : im - 1}$ and $W_{(i-1)m : im - 1}$ be denoted by $\hat{W}_i(t)$ and $W_i$ respectively. Similarly we denote $\hat{A}(t)_{S(i):}$ and $A_{S(i)}$ by $\hat{A}_i(t)$ and $A_i$ respectively. Then following identities hold,

$$A_iW_i = M_i$$

$$\hat{A}_i(t)\hat{W}_i(t) = \hat{M}_i(t)$$  \hspace{1cm} (41)

Note that $A_i$ has full-column rank. Let the left-inverse of $A_i$ be $A_i^\ast$. It is easy to see that,

$$\|A_i^\ast\|_{\infty,1} \leq \frac{m}{\rho_2}. $$  \hspace{1cm} (42)

From (41) we have,

$$\left( I + A_i^\ast(\hat{A}_i(t) - A_i) \right) \hat{W}_i(t) = W_i + A_i^\ast(\hat{M}_i(t) - M_i)$$

$$\implies \hat{W}_i(t) = \left( I + A_i^\ast(\hat{A}_i(t) - A_i) \right)^{-1} \left( W_i + A_i^\ast(\hat{M}_i(t) - M_i) \right)$$

$$\implies \hat{W}_i(t) = \left( I - A_i^\ast(\hat{A}_i(t) - A_i)(I + A_i^\ast(\hat{A}_i(t) - A_i)) \right) (W_i + A_i^\ast(\hat{M}_i(t) - M_i))$$

We can simplify further to yield,

$$\hat{W}_i(t) - W_i = A_i^\ast(\hat{M}_i(t) - M_i) - \left( A_i^\ast(\hat{A}_i(t) - A_i)W_i + \left( A_i^\ast(\hat{A}_i(t) - A_i) \right)^2 W_i \right)$$

$$- \left( A_i^\ast(\hat{A}_i(t) - A_i)A_i^\ast(\hat{M}_i(t) - M_i) + \left( A_i^\ast(\hat{A}_i(t) - A_i) \right)^2 A_i^\ast(\hat{M}_i(t) - M_i) \right)$$

Therefore by triangle inequality we have,

$$\|\hat{W}_i(t) - W_i\|_{\infty,1} = \|A_i^\ast(\hat{M}_i(t) - M_i)\|_{\infty,1} + \left\| \left( A_i^\ast(\hat{A}_i(t) - A_i)W_i + \left( A_i^\ast(\hat{A}_i(t) - A_i) \right)^2 W_i \right) \right\|_{\infty,1}$$

$$+ \left\| \left( A_i^\ast(\hat{A}_i(t) - A_i)A_i^\ast(\hat{M}_i(t) - M_i) + \left( A_i^\ast(\hat{A}_i(t) - A_i) \right)^2 A_i^\ast(\hat{M}_i(t) - M_i) \right) \right\|_{\infty,1}$$

Now we will bound each of the terms seperately as follows,

$$\|A_i^\ast(\hat{M}_i(t) - M_i)\|_{\infty,1} \leq \|A_i^\ast\|_{\infty,1} \|\hat{M}_i(t) - M_i\|_{\infty}$$

$$\leq \frac{m\epsilon_2}{\rho_2}.$$
Similarly we have,
\[ \left\| \left( \mathbf{A}_i^* (\hat{\mathbf{A}}_i(t) - \mathbf{A}_i) \mathbf{W}_i + \left( \mathbf{A}_i^* (\hat{\mathbf{A}}_i(t) - \mathbf{A}_i) \right)^2 \mathbf{W}_i \right) \right\|_{\infty,1} \leq \| \mathbf{A}_i^* \|_{\infty,1} (1 + \| \mathbf{A}_i^* \|_{\infty,1} \epsilon_1) \| \mathbf{W}_i \|_{\infty,\infty} \leq \frac{2m\epsilon_1}{\rho_2} \]

Finally the third term can be bounded as,
\[ \left\| \left( \mathbf{A}_i^* (\hat{\mathbf{A}}_i(t) - \mathbf{A}_i) \mathbf{A}_i^* (\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) + \left( \mathbf{A}_i^* (\hat{\mathbf{A}}_i(t) - \mathbf{A}_i) \right)^2 \mathbf{A}_i^* (\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) \right) \right\|_{\infty,1} \leq \left( \| \mathbf{A}_i^* \|_{\infty,1} \right)^2 \epsilon_1 \epsilon_2 + \left( \| \mathbf{A}_i^* \|_{\infty,1} \right)^3 \epsilon_1^2 \epsilon_2 \leq \frac{2m\epsilon_2}{\rho_2} \]

Therefore we have,
\[ \left\| \mathbf{W}_i(t) - \mathbf{W}_i \right\|_{\infty,1} \leq \frac{m(2\epsilon_1 + 3\epsilon_2)}{\rho_2} \]

We can repeat the same analysis for all \( i \in [l+1] \) to arrive at the required result. \( \square \)

### 5.9 Putting it together: Online Analysis

In this section we prove Theorem 8, which provides a parameter dependent upper bound on the regret of Algorithm 1 if \( \mathbf{W} \) and \( \mathbf{A} \) satisfy the \( \ell_1 \)-RRSVP and \( \ell_2 \)-RSVP. The regret bound provided here is in the parameter dependent regime, that is we assume a constant gap between the best arm and the rest for each context. More precisely let \( \Delta = \min_{s \in [t]} \left( u^*(s) - \max_{k \neq k^*(s)} \mathbf{U}_{sk} \right) \) be a fixed constant not scaling with \( L, K \) or \( t \). This falls under the purview of the random generative model because we allow for \( \Theta(K/m) \) deterministic rewards for each of the latent context. These conditions are expected to hold in real world data as each latent contexts are expected to have some unique arms which are significantly different from the others. In the said regime we reduce the regret bound of \( O(LK \log t) \) for general contextual bandit to only an \( O(L \text{poly}(m, \log K) \log t) \) dependence.

**Theorem 8.** In a contextual bandit setting suppose the reward matrix has the form \( \mathbf{U} = \mathbf{A} \mathbf{W} \) and each contexts \( s \) arrives independently with probability \( \beta_s \) for all \( s \in [L] \). Assume that \( L = \Omega(K \log K) \). If the problem parameters satisfy the following assumptions,

- \( \beta = \min_s \beta_s = \Omega(1/L) \).
- \( \mathbf{W} \in \mathbb{R}^{m \times K} \) satisfies \( \ell_1 \)-RRSVP with parameters \( (\delta, \rho_1, 2m') \)
- \( \mathbf{A} \in [0, 1]^{L \times m} \) satisfies \( \ell_2 \)-RSVP with parameters \( (\delta/L, \rho_2, 2m') \) and is separable [23].

then with probability atleast \( 1 - \delta \), Algorithm 1 with \( \epsilon_t = \min \left( 1, \frac{\theta(2m' + m)}{\beta t} \right) \) and \( \gamma(t) = \max \left( \frac{1}{t}, \frac{2}{\sqrt{\theta}} \right) \) has regret,

\[
R(T) \leq \frac{\theta(m + 2m') \log t}{\beta} + 4(L + K + 1)m' \log t + o(1) = O(L \text{poly}(m, m') \log t)
\]
where \( \theta \geq 4 \max \left( \frac{2m'(16+\Delta)\rho_1+32m}{\Delta \rho_1 \rho_2}, \frac{15}{\rho_1(1-\lambda)} \right)^2. \)

Before we proceed to the proof of our theorem, we need to introduce a few useful lemmas. The next lemma connects the chance of making an error in the exploit phase with the estimation errors in the system.

**Lemma 9.** Suppose at time \( t \), \( \| \hat{F}(t) - F \|_{\infty,\infty} \leq \epsilon_1(t) \) and \( \| \hat{M}_i(t) - M_i \|_{\infty} \leq \epsilon_2(t) \) for all \( i \in [l+1] \). If the following conditions hold,

\[
\begin{align*}
\epsilon_1(t) &\leq \min \left( \frac{\Delta \rho_1 \rho_2}{2m'(16+\Delta)\rho_2+32m}, \frac{\rho_1(1-\lambda)}{15} \right) \\
\epsilon_2(t) &\leq \frac{\Delta \rho_2}{12m} \\
E(t) &= 0
\end{align*}
\]

(43)

then \( k(t) = k^*(s_i) \), that is the optimal arm for the context is scheduled in the exploit phase.

**Proof.** If \( \epsilon_1(t) \leq \frac{\rho_1(1-\lambda)}{15} \), then by Lemma 6 we have,

\[
\| \hat{A}(t) - A \|_{\infty,1} \leq \frac{8m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)}
\]

(44)

Since we have,

\[
\begin{align*}
\epsilon_1(t) &\leq \frac{mp_1}{2m'(4\rho_2+m)} \\
\epsilon_2(t) &\leq \frac{\rho_2}{m}
\end{align*}
\]

it is easy to verify that the conditions of Lemma 8 are satisfied. Therefore we have,

\[
\| \hat{W}(t) - W(t) \|_{\infty,\infty} \leq \frac{m}{\rho_2} \left( \frac{16m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} + 3\epsilon_2(t) \right)
\]

(45)

Therefore we have,

\[
\begin{align*}
\| \hat{U}(t) - U \|_{\infty,\infty} &= \| AW - \hat{A}(t)W(t) \|_{\infty,\infty} \\
&\leq \| A \|_{\infty,1} \| W - \hat{W}(t) \|_{\infty,\infty} + \| A - \hat{A}(t) \|_{\infty,1} \| \hat{W}(t) \|_{\infty,\infty} \\
&\leq \frac{m}{\rho_2} \left( \frac{16m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} + 3\epsilon_2(t) \right) + \frac{8m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} \\
&\leq \frac{8m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} \left( 1 + \frac{2m}{\rho_2} \right) + 3\frac{m\epsilon_2(t)}{\rho_2}
\end{align*}
\]

Now, under the conditions of the lemma in (43), we have

\[
\frac{8m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} \left( 1 + \frac{2m}{\rho_2} \right) \leq \frac{\Delta}{4}
\]

\[
3\frac{m\epsilon_2(t)}{\rho_2} \leq \frac{\Delta}{4}
\]

25
This further implies that,
\[
\left\| \hat{U}(t) - U \right\|_{\infty, \infty} \leq \Delta/2
\]
This guarantees that we select the optimal arm at time-step \( t \).

The following lemma we prove that each entry of the matrices \( \hat{F}(t) \) and \( \hat{M}_i(t) \) for all \( i \in [l+1] \) are sampled sufficient number of times. Let \( T_{sj}(t) \) denote the number of samples obtained for the entry \( \hat{F}(t)_{sj} \). Similarly we define \( N^{(i)}(t)_{sj} \) as the number of sampled for the entry \( \hat{M}_i(t)_{sj} \).

**Lemma 10.** Suppose \( \epsilon_t = \frac{(m+2m')\theta}{m} \) where \( \beta = \min_s \beta_s \). Algorithm 1 ensures that,
\[
\mathbb{P}\left( T_{sj}(t) < \frac{\theta}{2} H_t \right) \leq \frac{1}{t^{\theta/12}}
\]
\[
\mathbb{P}\left( N^{(i)}(t)_{sj} < \frac{\theta}{2} H_t \right) \leq \frac{1}{t^{\theta/12}}
\]
and where \( H_n = \sum_{i=1}^{n} \frac{1}{i} \sim \log(n) \)

**Proof.** Let \( S_t \) denote the random variable describing the context at time \( t \). Let \( C_t \) denote the random variable denoting the the column of \( G(0) \) to be sampled provide \( E(t) = 1 \) and \( H_t = 1 \). Note that,
\[
\mathbb{E} [T_{sj}(t)] \geq \sum_{l=1}^{t} \mathbb{P}(S_l = s, E(l) = 1, H_l = 1, C_l = j)
\]
\[
\geq \sum_{l=1}^{t} \frac{\theta}{l} = \theta H_t
\]
Now, a straight forward application of Chernoff-Hoeffding’s inequality yields,
\[
\mathbb{P}(T_{sj}(t) < (1 - \delta)\mathbb{E} [T_{sj}(t)]) \leq \exp \left( -\frac{\delta^2}{3} \mathbb{E} [T_{sj}(t)] \right)
\]
\[
\leq \exp \left( -\frac{\delta^2}{3} \theta H_t \right)
\]
We can set \( \delta = 1/2 \) to get the required result. The same analysis works for \( N^{(i)}(t)_{sj} \). The corresponding entry is sampled if \( S_t = s_s(i) \). Let \( C'_t \) denote the column of \( G(i) \) to be sampled when \( E(t) = 1, S_t = s_s(i) \) and \( H_t = 0 \).
\[
\mathbb{E} [N^{(i)}(t)_{sj}] \geq \sum_{l=1}^{t} \mathbb{P}(E(t) = 1, S_t = s_s(i), H_t = 0, C'_l = j)
\]
\[
\geq \sum_{l=1}^{t} \frac{\theta}{l} = \theta H_t
\]
The same concentration inequality as before applies.

**Lemma 11.** Under the conditions of Lemma 10 we have,

\[
P\left( \left\| \hat{\mathbf{F}}(t) - \mathbf{F} \right\|_{\infty, \infty} > \epsilon_1(t) \right) \leq 4Lm' \exp \left( -\frac{\epsilon_1(t)^2 \theta \log t}{2} \right) + 2Lm' \frac{2}{t^{\theta/12}} \tag{46}
\]

**Proof.** The proof of this lemma is an application of Chernoff’s bound to the samples observed. Note that \( \mathbb{E} \left[ \hat{\mathbf{F}}(t) \right] = \mathbf{F} \). We have,

\[
P \left( |\hat{\mathbf{F}}(t)_{sj} - \mathbf{F}_{sj}| > \epsilon_1(t) \right) \leq P \left( |\hat{\mathbf{F}}(t)_{sj} - \mathbf{F}_{sj}| > \epsilon_1(t) \left| T_{sj}(t) \geq \frac{\theta}{2} H_t \right. \right) + P \left( T_{sj}(t) < \frac{\theta}{2} H_t \right)
\]

\[
\leq 2e^{-\frac{\epsilon_1(t)^2 \theta \log t}{2}} + \frac{1}{t^{\theta/12}}
\]

where the last inequality if due to lemma 10. Now, we can apply an union bound over all \( s \in [L] \) and \( j \in [m]' \) to obtain the required result. \( \square \)

Similarly we can bound the errors in estimating \( \mathbf{M}_t \)'s as in the lemma below.

**Lemma 12.** Under the conditions of Lemma 10 we have,

\[
P \left( \bigcup_{i \in \llbracket l + 1 \rrbracket} \left\{ \left\| \hat{\mathbf{M}}_t(t) - \mathbf{M}_t \right\|_{\infty, \infty} > \epsilon_2(t) \right\} \right) \leq 4(K+1)m' \exp \left( -\frac{\epsilon_2(t)^2 \theta \log t}{2} \right) + 2(K + 1)m' \frac{2}{t^{\theta/12}} \tag{47}
\]

**Proof.** The proof of this lemma is analogous to that of Lemma 11. We have the following chain,

\[
P \left( |\hat{\mathbf{M}}_t(t)_{sj} - \mathbf{M}_{t_{sj}}| > \epsilon_1(t) \right) \leq P \left( |\hat{\mathbf{M}}_t(t)_{sj} - \mathbf{M}_{t_{sj}}| > \epsilon_1(t) \left| T_{sj}(t) \geq \frac{\theta}{2} H_t \right. \right) + P \left( T_{sj}(t) < \frac{\theta}{2} H_t \right)
\]

\[
\leq 2e^{-\frac{\epsilon_2(t)^2 \theta \log t}{2}} + \frac{1}{t^{\theta/12}}
\]

We can apply union bound over all the entries of all the \( l + 1 \) matrices to get the result. \( \square \)

Now, we are at a position to prove our main theorem.

**Proof of Theorem 8.** We have \( \epsilon_t = \frac{(m+2m')\theta}{6t} \) where we set,

\[
\theta \geq 4 \max \left( \frac{2m'((16 + \lambda )\rho_2 + 32m)}{\Delta \rho_1 \rho_2}, \frac{15}{\rho_1(1 - \lambda)} \right)^2 \tag{48}
\]

By virtue of the \( l_1 \)-RRSVP property of \( \mathbf{W} \), the set \( S \) is \( \rho_1 \)-simplical with probability at least \( 1 - \delta \). Similarly, by Lemma 7 all the sets \( S(i) \) are good with probability at least \( 1 - \delta \). In what follows, we will assume that the above high probability conditions hold. Note that according to Lemmas 11 and 12 we have,

\[
P \left( \left\| \hat{\mathbf{F}}(t) - \mathbf{F} \right\|_{\infty, \infty} > \frac{2}{\sqrt{\theta}} \right) \leq 4Lm' \frac{2}{t} + o \left( \frac{1}{t^2} \right)
\]

\[
P \left( \bigcup_{i \in \llbracket l + 1 \rrbracket} \left\{ \left\| \hat{\mathbf{M}}_t(t) - \mathbf{M}_t \right\|_{\infty, \infty} > \frac{2}{\sqrt{\theta}} \right\} \right) \leq 4(K + 1)m' \frac{2}{t} + o \left( \frac{1}{t^2} \right) \tag{49}
\]

27
As \( U \in [0, 1]^{L \times K} \) the regret till time \( T \) can be bounded as follows,

\[
R(T) \leq \sum_{t=1}^{T} \mathbb{E}[1 \{ E(t) = 1 \}] + \sum_{t=1}^{T} \mathbb{E}[1 \{ E(t) = 0 \}] \mathbb{P}(k(t) \neq k^*(s_t)) \tag{50}
\]

By Lemma 9 we have that,

\[
\mathbb{P}(k(t) \neq k^*(s_t)) \leq \mathbb{P}\left( \left\| \hat{F}(t) - F \right\|_{\infty, \infty} > \frac{2}{\sqrt{\theta}} \right) + \mathbb{P}\left( \bigcup_{i \in [t+1]} \left\{ \left\| \hat{M}_i(t) - M_i \right\|_{\infty, \infty} > \frac{2}{\sqrt{\theta}} \right\} \right)
\]

We can combine this with (50) to get,

\[
R(T) \leq \frac{\theta(m + 2m') \log t}{\beta} + 4(L + K + 1)m' \log t + o(1) = O(L \text{poly}(m, m') \log t)
\]

if we assume that \( 1/\beta = O(L) \).

\[\square\]

### 5.10 Lower Bound for \( \alpha \)-consistent Policies

In this section we provide a problem dependent lower bound for the contextual bandit problem with latent contexts. The lower bound is established for a particular class of data-matrix \( U \) and for \( \alpha \)-consistent policies. For, any \( z_i \in \mathcal{Z} \) we define \( C(z_i) \) as,

\[
C(z_i) := \{ s \in \mathcal{S} : \alpha_{si} \neq 0 \}
\]

**Theorem 9.** Consider a problem instance \((U, A, W)\) such that \( \beta_s = 1/L \) for all \( s \in S \) and \( |C(z_i)| = L/m \) (assume that \( m \) divides \( L \)) for all \( z_i \in \mathcal{Z} \). Further, we assume that \( C(z_i) \cap C(z_j) = \emptyset \), for all \( z_i \neq z_j \). Then the regret of any \( \alpha \)-consistent policy is lower-bounded as follows,

\[
R(T) \geq (K - 1)mD(U) \left( (1 - \alpha)(\log(T/2m) - \log(L/m)) - \log(4KC) \right) \tag{51}
\]

for any \( T > \tau \), where \( C, \tau \) are universal constants independent of problem parameters and \( D(U) \) is a constant that depends on the entries of \( U \) and is independent of \( L, K \) and \( m \).

In order to prove Theorem 9 we introduce an inequality from the hypothesis testing literature.

**Lemma 13 ([27]).** Consider two probability measures \( P \) and \( Q \), both absolutely continuous with respect to a given measure. Then for any event \( \mathcal{A} \) we have:

\[
P(\mathcal{A}) + Q(\mathcal{A}^c) \geq \frac{1}{2} \exp\{-\min(\text{KL}(P||Q), \text{KL}(Q||P))\}
\]
Proof of Theorem 9. Note that the conditions in the theorem imply that there are \( m \) distinct latent contexts and there are \( L/m - 1 \) copies for each of them. For any \( z_i \in Z \) let us define \( T(z_i) = \sum_{t=1}^{T} \mathbb{1}\{S_t \in C(z_i)\} \). With some abuse of notation we also define \( k^*(z_i) \) as the index of the optimal arm and \( \Delta(z_i) \) as the gap between the optimal and second optimal arm for all contexts in \( C(z_i) \). By the assumptions in the theorem we have,

\[
\mathbb{E}[T(z_i)] = \frac{T}{m}
\]

Let \( E_i \) be the event \( \{\frac{T}{2m} \leq T(z_i) \leq \frac{3T}{m}\} \). Let \( E^c = \{\bigcup_{z_i \in Z} E^c_i\} \). By a simple application of Chernoff bound we have,

\[
P(\bigcup_{z_i \in Z} E^c_i) \leq 2m e^{-T/12} = o\left(\frac{1}{T^2}\right)
\]

Fix a \( z_i \in Z \) and let \( k \) be the index of an arm that is not optimal for any of the contexts that belong to \( C(z_i) \). Let us create another system with parameter \( (U', A, W') \) where we make the entry \( W_{sk} = \lambda = \frac{U_{max} + 1}{2} \) where \( U_{max} = \max_{s,k} U_{sk} \), while everything else remains the same including the coefficients of the convex combinations relating the observed contexts to the latent contexts. Note that this implies that in the second system arm \( k \) is optimal for all \( s \in C(z_i) \). Let \( A \) be the event defined as follows,

\[
A := \left\{ \sum_{\{t: S_t \in C(z_i)\}} \mathbb{1}\{X_t = k\} \geq \frac{T(z_i)}{2} \right\}
\]

Now, in the system with parameter \( U \) for any \( s \in C(z_i) \) we have,

\[
\mathbb{E}\left[\sum_{\{t: S_t = s\}} \mathbb{1}\{X_t = k\}\right] \leq C T(s)^\alpha
\]

if \( T(s) \geq \tau \), since the policy in consideration is \( \alpha \)-consistent. Here, \( \tau, C \) are universal constants. By an application of Jensen’s inequality we have,

\[
\mathbb{E}\left[\sum_{\{t: S_t \in C(z_i)\}} \mathbb{1}\{X_t = k\}\right] \leq C |C(z_i)|^{1-\alpha} T(z_i)^\alpha
\]

Let \( \mathbb{P}_U^T \) and \( \mathbb{P}'_U^T \) be the distributions corresponding to the chosen arms and rewards obtained for \( T \) plays for the two instances under a fixed \( \alpha \)-consistent policy. Now we can apply Markov’s inequality to conclude that,

\[
\mathbb{P}_U(A) \leq \frac{2 C |C(z_i)|^{1-\alpha}}{T(z_i)^{1-\alpha}}
\]

\[
\mathbb{P}'_U(A^c) \leq \frac{2(K-1) C |C(z_i)|^{1-\alpha}}{T(z_i)^{1-\alpha}}
\]  

(52)
Now from Lemma 13 we have,

\[
\text{KL}\left(\mathbb{P}_U^T, \mathbb{P}_{U'}^T\right) \geq (1 - \alpha) (\log(T(z_i)) - \log(L/m)) - \log(4KC)
\] (53)

Using standard methods from the bandit literature it can be shown that,

\[
\text{KL}\left(\mathbb{P}_U^T, \mathbb{P}_{U'}^T\right) = \sum_{s \in \mathcal{C}(z_i)} \sum_{t : S_t = s} \text{KL} (U_{sk}, \lambda) \mathbb{E}_U [\mathbf{1}\{X_t = k\}]
\]

Let us define the regret incurred during the time-steps where \(S_t \in \mathcal{C}(z_i)\) as \(R(T(z_i))\). We can follow the same procedure for all the sub-optimal arms which yields the following bound,

\[
R(T(z_i)) \geq \Delta(z_i) \sum_{k \neq k^*(z_i)} \sum_{s \in \mathcal{C}(z_i)} \sum_{t : S_t = s} \mathbb{E}_U [\mathbf{1}\{X_t = k\}]
\]

\[
\geq \left( \arg\min_k \frac{(K - 1)\Delta(z_i)}{\text{KL}(U_{sk}, \lambda)} \right) ((1 - \alpha) (\log(T(z_i)) - \log(L/m)) - \log(4KC))
\]

Let \(D(U) = \left( \arg\min_{z_i, k} \frac{(K-1)\Delta(z_i)}{\text{KL}(U_{sk}, \lambda)} \right)\). Now, we have

\[
R(T) = \sum_{z \in \mathcal{Z}} \mathbb{E} [R(T(z_i))]
\]

\[
\geq D(U)(K - 1)\mathbb{E} \sum_{z \in \mathcal{Z}} ((1 - \alpha) (\log(T(z_i)) - \log(L/m)) - \log(4KC))
\]

Now, using the fact that \(T(z_i) \geq \frac{T}{2m}\) given \(E\), we have

\[
R(T) = \sum_{z \in \mathcal{Z}} \mathbb{E} [R(T(z_i))] = \sum_{z \in \mathcal{Z}} \mathbb{E} [R(T(z_i))|E] \mathbb{P}(E) + \mathbb{E} [R(T(z_i))|E^c] \mathbb{P}(E^c)
\]

\[
\geq D(U)(K - 1)m ((1 - \alpha) (\log(T/2m) - \log(L/m)) - \log(4KC)) + o(1)
\]

\[\square\]

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In Advances in Neural Information Processing Systems, pages 2312–2320, 2011.

[2] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. arXiv preprint arXiv:1209.3352, 2012.

[3] S. Arora, R. Ge, and A. Moitra. Learning topic models—going beyond svd. In Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, pages 1–10. IEEE, 2012.

[4] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In Advances in Neural Information Processing Systems, pages 1342–1350, 2015.
[5] R. Bell, Y. Koren, and C. Volinsky. The bellkor solution to the netflix prize, 2007.

[6] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721, 2012.

[7] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In International Conference on Artificial Intelligence and Statistics, pages 208–214, 2011.

[8] A. Damle and Y. Sun. Random projections for non-negative matrix factorization. arXiv preprint arXiv:1405.4275, 2014.

[9] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. Neural computation, 21(3):793–830, 2009.

[10] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In Advances in Neural Information Processing Systems, pages 586–594, 2010.

[11] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing, volume 1. Springer, 2013.

[12] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. arXiv preprint arXiv:1401.8257, 2014.

[13] N. Gillis. The why and how of nonnegative matrix factorization. arXiv preprint arXiv:1401.5226, 2014.

[14] N. Gillis and Stephen A V. Fast and robust recursive algorithms for separable nonnegative matrix factorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36(4):698–714, 2014.

[15] N. Guan, D. Tao, Z. Luo, and B. Yuan. Online nonnegative matrix factorization with robust stochastic approximation. Neural Networks and Learning Systems, IEEE Transactions on, 23(7):1087–1099, 2012.

[16] M. Hardt and M. Wootters. Fast matrix completion without the condition number. arXiv preprint arXiv:1407.4070, 2014.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American statistical association, 58(301):13–30, 1963.

[18] S. Jukna. Extremal combinatorics: with applications in computer science. Springer Science & Business Media, 2011.

[19] E. Kaufmann, O. Cappe, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In International Conference on Artificial Intelligence and Statistics, pages 592–600, 2012.
[20] J. Kawale, H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla. Efficient thompson sampling for online matrix-factorization recommendations. In Advances in Neural Information Processing Systems, pages 1297–1305, 2015.

[21] O. Maillard and S. Mannor. Latent bandits. In Proceedings of The 31st International Conference on Machine Learning, pages 136–144, 2014.

[22] J. Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, 2009.

[23] B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In Advances in Neural Information Processing Systems, pages 1214–1222, 2012.

[24] A. Salomon, J. Audibert, and I. Alaoui. Regret lower bounds and extended upper confidence bounds policies in stochastic multi-armed bandit problem. arXiv preprint arXiv:1112.3827, 2011.

[25] A. Slivkins. Contextual bandits with similarity information. The Journal of Machine Learning Research, 15(1):2533–2568, 2014.

[26] K. Stromberg. Probability for analysts. CRC Press, 1994.

[27] A. Tsybakov. Introduction to nonparametric estimation. Springer Science & Business Media, 2008.

[28] S. Vavasis. On the complexity of nonnegative matrix factorization. SIAM Journal on Optimization, 20(3):1364–1377, 2009.

[29] C. Wang, S. Kulkarni, and V. Poor. Arbitrary side observations in bandit problems. Advances in Applied Mathematics, 34(4):903–938, 2005.

[30] H. Wu, R Srikant, X. Liu, and C. Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In Advances in Neural Information Processing Systems, pages 433–441, 2015.

[31] Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. arXiv preprint arXiv:1604.06743, 2016.