Contextual Bandits with Latent Confounders: An NMF Approach

Rajat Sen
The University of Texas at Austin

Karthikeyan Shanmugam
IBM Thomas J. Watson Research Center

Murat Kocaoglu
The University of Texas at Austin

Alexandros G. Dimakis
The University of Texas at Austin

Sanjay Shakkottai
The University of Texas at Austin

Abstract

Motivated by online recommendation and advertising systems, we consider a causal model for stochastic contextual bandits with a latent low-dimensional confounder. In our model, there are L *observed contexts* and K *arms* of the bandit. The observed context influences the reward obtained through a *latent* confounder variable with cardinality m ($m \ll L, K$). The arm choice and the latent confounder causally determines the reward while the observed context is correlated with the confounder. Under this model, the $L \times K$ mean reward matrix $\mathbf{U}$ (for each context in $[L]$ and each arm in $[K]$) factorizes into non-negative factors $\mathbf{A}$ ($L \times m$) and $\mathbf{W}$ ($m \times K$). This insight enables us to propose an ϵ-greedy NMF-Bandit algorithm that designs a sequence of *interventions* (selecting specific arms), that achieves a balance between learning this low-dimensional structure and selecting the best arm to minimize *regret*. Our algorithm achieves a regret of $\mathcal{O}\left(L\text{poly}(m, \log K)\log T\right)$ at time T, as compared to $\mathcal{O}(LK\log T)$ for conventional contextual bandits, assuming a constant gap between the best arm and the rest for each context. These guarantees are obtained under mild sufficiency conditions on the factors that are weaker versions of the well-known Statistical RIP condition. We further propose a class of generative models that satisfy our sufficient conditions, and derive a lower bound of $\mathcal{O}\left(Km\log T\right)$. These are the first regret guarantees for online matrix completion with bandit feedback, when the rank is greater than one. We further compare the performance of our algorithm with the state of the art, on synthetic and real world data-sets.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

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 meet their interests and simultaneously 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* [9].

The classic K-armed bandit problem has been studied extensively for 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 [9] scales as $O(K\log T)$ when there is a constant gap between the best arm and the rest.

When side-information is available, a popular model is the contextual bandit, where the side information is encoded through *observed contexts*. In the stochastic setting, at each time an observed context $s \in [L]$ is revealed, and the observed context influences the reward statistics of the K arms. Thus, there are $(K \times L)$ reward parameters $\{U_{sk}\}$ (encoded through the reward matrix $\mathbf{U}$) that need to be learned, one per each arm and observed context. Since there are $(K \times L)$ reward parameters, it has been shown [9, 40] that the best expected regret obtainable scales as $O(KL \log T)$.

Netflix Example: Consider the task of recommending movies to user profiles on Netflix. A user profile along with the past browsing history, social and demographic information is the *observed context*. The list of movies that can be recommended to any user are the arms of the bandit. In this setting with millions of users and items, standard contextual bandit algorithms are rendered impractical due to the $K \times L$ scaling.

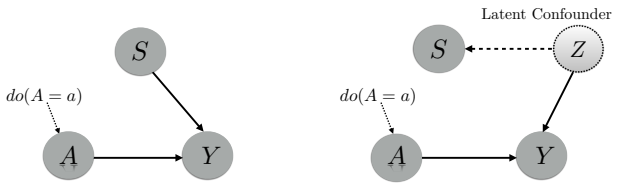
Therefore, it is important to exploit that in most practical situations, the underlying factors affecting the rewards may have a low-dimensional structure. Although this low dimensional structure is often not observable (latent), we will show that it can be leveraged to obtain better regret bounds. In the context of Netflix, there are millions of user profiles but the preference of users towards an item may be represented by a combination of only a handful of *moods*, where these *moods* lie in a much lower dimension. This is further corroborated by the fact that the Netflix data-set, which has more than 100 million movie ratings, can be approximated surprisingly well by matrices of rank as low as 20 [7]. Crucially however, these *moods* cannot be directly observed by a learning algorithm.

This problem of a contextual bandit with a latent structure has direct analogy with problem of designing structural *interventions* (forcing variables to take particular values) in causal graphs, a class of problems that is of increasing importance in social sciences, economics, epidemiology and computational advertising [32, 8].

A Causal Perspective: A *causal model* [32] is a directed graph that encodes causal relationships between a set of random variables, where each variable is represented by a node of the graph (see Figure 1a). This example has a directed graph with 3 variables, where the variable Y has two parents $\{S, A\}$.

To illustrate the connection between contextual bandits and causal models, consider again the Netflix example, which can be mapped to the causal graph in Figure 1a. Here, the *reward* Y (satisfaction of the user) is causally dependent on two quantities – the *observed context* (user profile in Netflix) described by S, and the *arm selection* (the recommended movie) described by the variable A. Setting A to a particular value is equivalent to playing a particular arm (act of recommending an item). In this example, A is the *only* variable that can be directly controlled by the algorithm; in the language of causality this is known as an *intervention* [32] denoted by $do(A = a)$.

More specifically, this contextual bandit setting maps to the causal graph problem of affecting a target variable Y (satisfaction of users), through *limited interventional capacity* (only being able to recommend a movie) when other observable causes (user profiles and contextual information) affecting the target variable are present but cannot be controlled. This is precisely the model in Figure 1a. An identical structural equation model has been defined in Figure 8 of [8].



(a) Causal graph representing regular contextual bandits

(b) Causal graph representing contextual bandits with latent confounders.

Figure 1: Comparison between regular contextual bandits and contextual bandits with latent confounders through causal graphs.

Latent Confounders: In this causal framework, it is possible to formally capture the implications of latent effects, such as the *moods* in the context of Netflix. Consider the modified causal model in Figure 1b. The new variable Z denotes a *latent confounder* (mood) that is causally connected to the *observed context* and also causally affects the reward Y. The latent confounder Z takes values in $\{1, 2, ..., m\}$, where $m \ll L, K$.

The goal here is to develop an efficient algorithm that chooses the sequence of *limited interventions* (i.e. a sequence of $do(A = a)$ actions) to achieve a balance between learning this latent variable (indirectly learning Z) from observed rewards, and maximizing the observed reward under the given (but not intervenable) observed context S.

In the setting of *contextual bandits* with L observed contexts and K arms, we note that the presence of the m-dimensional latent confounder leads to a factorization of the $L \times K$ reward matrix $\mathbf{U}$ into non-negative factors $\mathbf{A}$ (an $L \times m$ matrix) and $\mathbf{W}$ (a $m \times K$ matrix). We leverage this latent low-dimensional structure to develop an ϵ-greedy NMF-Bandit algorithm that achieves a balance between learning the hidden low-dimensional

structure (indirectly learning Z), and selecting the best arm to minimize regret. In the setting of *causality*, this result thus demonstrates an approach to designing a sequence of *interventions* with *limited capacity* to control a reward variable, in the presence of other (possibly latent) variables affecting the reward that *cannot* be intervened upon.

1.1 Main Contributions

The main contributions of this paper are as follows:

1. (Model for Latent Confounding Contexts) We investigate a causal model for contextual bandits (Figure 1b), which, compared to the conventional model, allows more degrees of freedom through the unobservable context variable. This allows us to better capture real-world scenarios. In particular, our model has *(a) Latent Confounders* representing unobserved low-dimensional variables affecting the mean rewards of the bandit arms under an observed context; and *(b) Limited Interventional Capacity* signifying that the observed contexts (eg. user profiles) *cannot* be intervened upon.

In the contextual bandit setting with L observed contexts and K arms, this translates into a decomposition of the $L \times K$ reward matrix $\mathbf{U} = \mathbf{AW}$, where $\mathbf{A}$ (non-negative $L \times m$ matrix) represents the relation between X (observed contexts) and Z (hidden confounder), while $\mathbf{W}$ (non-negative $m \times K$ matrix) encodes the relation between Y(reward) and Z.

2. (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 $\mathbf{U}$.

A direct application of results from existing noisy matrix completion literature is infeasible in the bandit setting. In the literature, 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 [20]. In the bandit setting where errors occur due to sampling, this would lead to a regret of at least would lead to $O(LK \log T)$ in the presence of sampling errors. We provide further insights in Section A.2 in the appendix.

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 $\mathbf{A}$ and $\mathbf{W}$ satisfy conditions weaker than the well-known statistical RIP property [36]. Further, we prove a lower bound for this setting which is only $\text{poly}(m, \log K)$ factors away from our upper bound. This the the first work which has provable guarantees for matrix completion with bandit feedback for rank greater than one.

3. (Generative Models for A and W) We propose a family of generative models for the factors $\mathbf{A}$ and $\mathbf{W}$ 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 [16]).

Finally, we numerically compare our algorithm with contextual versions of UCB-1, Thompson Sampling algorithms [9] and online matrix factorization algorithms [25] on synthetic and real-world data-sets.

1.2 Related Work

The current work falls at the intersection of learning of low-dimensional causal structures and multi-armed bandit problems. We briefly review the areas of literature 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, the best known regret bounds scale as $O(\sqrt{LKT \log K})$ [9, 39] 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 $O(LK \log T)$ [40]. Contextual bandits with linear payoff functions have been analyzed in [2, 12] in the adversarial setting, while in [1] it has been analyzed in the stochastic setting. In [15] the authors have expanded this model for the generalized linear model regime.

However, these models require one of the low-dimensional features to be known a priori, while our algorithm learns both the features from sampled data. Another related line of work is in the online clustering of bandits [17, 31, 30]. In this framework, the features of the arms can be directly observed, which is the fundamental difference from our paper.

Causality and Bandits: Recently, contextual bandit algorithms have found use within the framework of causality. In [5], the authors investigate a similar latent confounder model. However [5] does not consider our scaling regime nor provide theoretical guarantees (and has a very different algorithm).

In [29], a causal model for observing feedback has been introduced in the best arm identification regime. However, in their model all the variables can be intervened upon. Moreover, the states of all the non-intervened variable including the reward is revealed after the intervention is made. In this work, we focus on a more realistic case where only some of the variables can be intervened upon and in fact some of the variables cannot be observed directly. Further, side information about the observed variables are revealed before an intervention has to be made. The reward is the only extra information that is revealed after each intervention.

Online Matrix factorization : The non-negative matrix factorization (NMF) problem has generated a lot of interest in the area of semi-supervised topic modeling. Arora et al. have shown that if the matrix is separable and has some robustness properties [4], then NMF is solvable efficiently. Since then, there has been a lot of work in proposing efficient scalable algorithms for NMF, out of which [18, 13, 33] are of particular interest. There has been some progress in online NMF [14, 19] which aims to update the features efficiently in a streaming sense. To the best of our knowledge there has been no work in NMF with bandit feedback with theoretical guarantees. [27, 25] propose algorithms for online matrix factorization, however they only have theoretical analysis for the *rank* 1 case.

2 Problem Statement and Results

2.1 System Model

Observed Contexts and Latent Confounders: We consider a stochastic bandit model represented by the causal graph in Figure 1b. The variable S denoting the observed context takes values in $\mathcal{S} = \{1, 2, \cdots, L\}$, while the variable A determines the arm that has been pulled taking values in $\mathcal{A} = \{1, 2, \cdots, K\}$. The variable Z denotes the latent confounding contexts and takes values in $\mathcal{Z} = \{z_1, z_2 \cdots z_m\} \subset \mathcal{S}$, where $m \ll L, K$. The causal model results in the bayesian factorization of the joint distribution of S, Y and Z. A natural interpretation is that, at any time nature chooses a latent context $z \in \mathcal{Z}$, and based on that, a context $s \in \mathcal{S}$ is actually observed. We denote the posterior probability of a *latent* context z given an observed context s as,

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

Let $\mathbf{A}$ be the matrix with elements α_{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 $\mathbf{I}_{m \times m}$. This is essentially the well-known *separability* condition [33]. We also define the marginal probability of observing a context $s \in S$ as $\mathbb{P}(S = s) = \beta_s$, $\forall s \in \mathcal{S}$. 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 $S_t = s_t \in \mathcal{S}$; *(ii)* After observing the context the algorithm selects an arm $A_t = a_t \in \mathcal{A}$ which is the *intervention* $do(A = a_t)$; and *(iii)* The algorithm then obtains a Bernoulli reward Y_t with mean U_{s_t, a_t}. The mean rewards U_{s_t, a_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 $\mathbf{W}$. W_{ik} specifies the mean reward for arm k when the latent context is z_i. For all observed contexts $s \in \mathcal{S}$, the mean rewards are given by the matrix $\mathbf{U}$. 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 $\mathbf{U} = \mathbf{AW}$. Since the latent contexts $\mathcal{Z}$ are also a subset of observed contexts, the matrix $\mathbf{A}$ contain a $\mathbf{I}_{m \times m}$ sub-matrix. This is equivalent to the separability condition and is widely used in the NMF literature (see [18]). $\mathbf{A}$ represent the relation $S \longleftrightarrow Z$ while the matrix $\mathbf{W}$ denotes the relation $Z \longrightarrow Y$ in the causal model of Figure 1b.

Regret: The goal is to minimize regret (also known as pseudo-regret [9]) when compared to a *genie* strategy which knows the matrix $\mathbf{U}$. Let us denote the best arm under a context $s \in \mathcal{S}$ 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,

$$R(T) = \sum_{s \in \mathcal{S}} \sum_{\{t \in [T] : S_t = s\}} (u^*(s) - \mathbb{E}[Y_t]) \qquad (1)$$

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 $\mathbf{U}$ and $\mathcal{Z}$, however we assume that m (the latent dimension) is a known scalar parameter. We work in the problem dependent setting, where there is a gap (bounded away from zero) between the mean reward of the best arm and the second best for every observed context. Let the gap (Δ), be defined as, $\Delta = \min_{s \in [L]} \min_{k \neq k^*(s)} u^*(s) - U_{sk}$.

2.2 Notation

We denote matrices by bold capital letters (e.g. $\mathbf{U}$) and vectors with bold small letters (e.g. $\mathbf{x}$). For an $L \times K$ matrix $\mathbf{U}_{S,:}$ denotes the sub-matrix restricted to the rows in $S \subset [L]$, while $\mathbf{U}_{:,R}$ denotes the sub-matrix restricted to the columns in $R \subset [K]$. $\sigma_m(\mathbf{P})$

denotes the m-th smallest singular value of $\mathbf{P}$. $\|\mathbf{x}\|_p$ denotes the ℓ_p-norm of $\mathbf{x}$. For, a matrix $\|\mathbf{U}\|_{\infty,1}$ refers to the maximum ℓ_1-norm among all the rows while $\|\mathbf{U}\|_2$ and $\|\mathbf{U}\|_F$ denotes its spectral and Frobenius norms respectively. $\|\mathbf{U}\|_{\infty,\infty}$ denotes the maximum absolute value of an element in the matrix. Ber(p) denotes a Bernoulli random variable with mean p.

2.3 Main results

We first provide few definitions before presenting our main results.

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

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

In our work, we require the matrices ($\mathbf{W}$ and $\mathbf{A}$) to satisfy some weaker versions of the 'statistical RIP property' (RIP - restricted isometry property). This property has been well studied in the sparse recovery literature [6, 11, 37, 36, 10]. Statistical RIP property is a randomized variant of the well-known RIP condition [16]. RIP requires the extreme singular values to be bounded for sub-dictionaries formed by *any* k columns (or rows) of a dictionary for a suitable k. Statistical RIP property is a weaker probabilistic version where extreme singular values need to be bounded for random sub-dictionaries with high probability when k random columns are chosen out of a dictionary to form the random subdictionary. We note that this same property goes by different names such as weak RIP property [11] and quasi-isometry property [10] in the literature. The terminology we adopt in this work is from [6].

Definition 3. (Statistical RIP Property - StRIP) *An $L \times m$ matrix ($L \geq m$) P, whose rows have unit ℓ_2 norm, satisfies the ℓ_2-Statistical RIP Property (ℓ_2-StRIP) with constants (ϵ, ρ, m'), if*

$$\mathrm{Pr}_{|S|=m'}(1-\rho \leq \sigma_{\min}(\mathbf{P}_{S,:}) \leq \sigma_{\max}(\mathbf{P}_{S,:}) \leq 1+\rho) \geq 1-\epsilon,$$

where the probability is taken over sampling a set S of size m' uniformly from $[L]$.

In our work, we only need a weaker version of StRIP condition to hold. We only need that the smallest singular value be bounded below for random sub-matrices and we work with un-normalized matrices. Hence, we have the following version which we will use:

Definition 4. (ℓ_2 Weak Statistical RIP Property - ℓ_2-WStRIP) *An $L \times m$ matrix ($L \geq m$) P satisfies the ℓ_2-Weak Statistical RIP Property (ℓ_2-WStRIP) with constants (ϵ, ρ, m') if* $\mathrm{Pr}_{|S|=m'}(\sigma_{\min}(\mathbf{P}_{S,:}) \geq \rho) \geq 1-\epsilon$ *where the probability is taken over sampling a set S of size m' uniformly from $[L]$.*

For one of the matrices among $\mathbf{W}$ and $\mathbf{A}$, we need its random sub-matrices to satisfy weaker RIP-like conditions in the ℓ_1 sense.

Definition 5. (ℓ_1 Weak Statistical RIP Property - ℓ_1-WStRIP) *An $m \times K$ matrix ($K \geq m$) P satisfies the ℓ_1-weak statistical RIP property (ℓ_1-WStRIP) with constants (ϵ, ρ, m') if* $\mathrm{Pr}_{|S|=m'}(\psi^1_m(\mathbf{P}_{:,S}) \geq \rho) \geq 1-\epsilon$ *where the probability is taken over sampling a set S of size m' uniformly from $[K]$.*

In what follows, we assume that $\mathbf{W}$ satisfies ℓ_1-WStRIP and $\mathbf{A}$ satisfies ℓ_2-WStRIP. Note that in Section 2.4 we provide reasonable generative models for $\mathbf{W}$ and $\mathbf{A}$ that satisfy these conditions with high probability.

Now we are at a position to state Theorem 1 which shows 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 $\mathbf{U} = \mathbf{AW}$. Suppose $\mathbf{A}$ is separable [33]. Let $\mathbf{A}$ satisfy ℓ_2-WStRIP with constants $(\delta/L, \rho_2, m'_1)$ while $\mathbf{W}$ satisfies ℓ_1-WStRIP with constants (δ, ρ_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,*

$$R(T) = O\left(L\frac{\mathrm{poly}(m, \log(K))}{\Delta^2}\log(T)\right) \qquad (2)$$

with probability at least $1-\delta$. Here, $\mathrm{poly}(m, \log(K)) = O\left(m^5 \log^2 K\right)$.

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. It should be noted that in practice our algorithm has much lesser regret than $O(Lm^5 \log T)$. This can be observed in Section 4, where our algorithm performs well even if we set the *explore* rate much lower than what is prescribed.

Remark: In prior works [6, 11, 37, 36, 10] the statistical RIP property was established by relating it to the incoherence parameter μ of a matrix $\mathbf{B}$ which is defined as $\mu(\mathbf{B}) = \max_{i\neq j}|\mathbf{b}_i^T\mathbf{b}_j|$. In some works, the average of these incoherence parameters has been used instead. We note that matrices $\mathbf{A}$ and $\mathbf{W}$ are non-negative. Hence, directly using analysis based on controlling dot-products among rows and columns is not useful in this scenario. Hence, we propose generative models for $\mathbf{A}$ and $\mathbf{W}$ that satisfy the properties listed above with high probability even when they are not incoherent. We

also explain why these generative models are extremely reasonable for our setting.

2.4 Generative Models for W and A

We briefly describe our semi-random generative models for $\mathbf{W}$ and $\mathbf{A}$ that satisfy the weak statistical RIP conditions. We refer to Section A.3 for a more detailed discussion of the generative models.

1. *Random+Deterministic Composition*: A significant fraction of entries of $\mathbf{W}$ and $\mathbf{A}$ are *arbitrarily deterministic*. $O(1/m)$ fraction of columns of $\mathbf{W}$ and $O(1)$ fraction of rows of $\mathbf{A}$ are deterministic. In addition, we assume that a sub-matrix in the deterministic part of $\mathbf{A}$ is an identity matrix to account for the separability condition [33]. The rest of the entries are mean shifted, bounded sub-gaussian random variables with some additional mild conditions. Uniform prior on reward that has been used in bandit setting [26] reduces to a special case of this model.

2. *Bounded randomness in the random part:* The random entries of both $\mathbf{W}$ and $\mathbf{A}$ are in "general position", i.e., they arise from mean shifted bounded sub-gaussian distributions (see Section A.3, and also [16] for similar conditions in compressed sensing literature). The mean shifts in the random parts of $\mathbf{A}$ and $\mathbf{W}$, the support of the sub-gaussian randomness satisfy technical conditions to make sure that row sum of $\mathbf{A}$ is 1 and to ensure that the weak statistical RIP conditions are satisfied.

One of our main results is stated as Theorem 2, which implies that if $\mathbf{W}$ comes from our generative model then with high probability projecting it onto a small random subset of its columns preserves the α-robust simplical property [33] which is a key step in our algorithm.

Theorem 2. *Let* $m' \geq \frac{512}{21\tilde{c}} m \log(eK)$. *Let* $\mathbf{W}$ *follow the random model in Section A.3.* $\mathbf{W}$ *satisfies* (ℓ_1*-WStRIP) with constants* $(2\exp(-c_1 \log(eK)), \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}}, 2m')$ *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* $\mathbf{W}$.

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

Theorem 3. *Let* $m' \geq \frac{512}{21\tilde{c}} m \log(eL)$. *Let* $\mathbf{A}$ *follow the random model in Section A.3.* $\mathbf{A}$ *satisfies* (ℓ_2*-WStRIP) with constants* $(2\exp(-c_2 m \log(eL)), \frac{1}{20}\frac{\sqrt{m'}}{m}, 2m')$ *with probability at least* $1-\exp(-c_2' m \log(eL))$. *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* $\mathbf{A}$.

The proof of these theorems are available in the appendix in Section A.4.

2.5 Lower Bound

We prove a problem-specific regret lower bound for a specific class of parameters $(\mathbf{U}, \mathbf{W}, \mathbf{A})$ which is only a $\text{poly}(m, \log(K))/\Delta$ factor away from the upper bound achieved by our algorithm. The lower bound holds for all policies in the class of α-consistent policies [34] defined below.

Definition 6. *A scheduling policy is said to be* α*-consistent if given any problem instance* $\mathbf{U}$ *we have,* $\mathbb{E}\left[\sum_{\{t\in[T]:S_t=s\}} \mathbb{1}\{X_t = k\}\right] = O(T(s)^\alpha)$ *for all* $k \neq k^*(s)$, $s \in \mathcal{S}$, *where* $\alpha \in (0,1)$, $T(s) = \sum_{t=1}^{T} \mathbb{1}\{S_t = s\}$

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

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

for any $T > \tau$, *where* C, τ *are universal constants independent of problem parameters and* $D(\mathbf{U}) = O(1/\Delta)$ *is a constant that depends on the entries of* $\mathbf{U}$ *and is independent of* L, K *and* m.

The proof of this theorem has been deferred to the appendix in Section A.11 where we specify the class of problem parameters for which we construct this bound.

3 NMF-Bandit Algorithm

In this section we present an ϵ-greedy algorithm that we call NMF-Bandit algorithm. Our algorithm takes advantage of the the low-dimensional structure of the reward matrix. The algorithm *explores* with probability ϵ_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. A detailed pseudo-code of our algorithm has been presented as Algorithm 1 in the appendix. The key steps in the algorithm are:

(a) At each time t and with probability ϵ_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 α it samples a random arm from a subset $S \subset [K]$

of arms. $|S| = 2m'$ for $m' = O(m \log(K))$. The set S is a randomly chosen at the onset and kept fixed there after. This is Step 6 of Algorithm 1.

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. This is Step 7 of Algorithm 1.

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

Step 3 – (Choose best arm for current observed context): Compute estimate $\hat{\mathbf{A}}(t)$ as detailed in Step 10 of Algorithm 1, using Hottopix. Estimate $\hat{\mathbf{W}}(t)$ as detailed in Step 11 of Algorithm 1. Let $\hat{\mathbf{U}}(t) = \hat{\mathbf{A}}(t)\hat{\mathbf{W}}(t)$. The algorithm plays the arm given by $\arg\max_{k\in[K]} \hat{\mathbf{U}}(t)_{s_t,k}$, i.e., the best arm for the observed context according to current estimates.

For solving the NMF to obtain $\hat{\mathbf{A}}(t)$, we use a robust version of Hottopix [33, 18] as a sub-routine. 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 A.1. 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.

A more detailed version of our main result (Theorem 1) has been provided in (Theorem 8) in the appendix, along with a detailed proof. Theorem 8 exactly specifies the algorithm parameters ϵ_t, α and m' under which we obtain the regret guarantees. We provide some key theoretical insights and a brief proof sketch in Section A.2 in the appendix. In particular we discuss in detail why usual matrix completion techniques would fail to provide regret guarantees that are $o(KL\log(T))$. We explain the challenges of dealing with sampling noise and how we overcome them through careful design of the arms to *explore*.

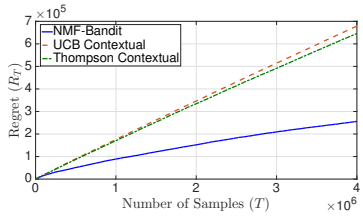
4 Empirical Validation

We validate the performance of our algorithm against various benchmarks on real and synthetic datasets. We compare our algorithms against contextual versions of UCB-1 [9] and Thompson sampling [3]. To be more precise, these algorithms proceed by treating each context separately and applying the usual K-armed version of the algorithms to each context. We also compare the performance of our algorithm to this recent algorithm [25] for stochastic rank 1 bandits. In [25] the problem setting is different. Therefore, whenever we compare the performance with this algorithm the experiments have been performed in the setting of [25], which we call **S2**. The more realistic setting of our paper will be denoted by **S1**. The two settings are, (i) **S1** : The arrival of the contexts *cannot* be controlled by the algorithm and the regret is w.r.t the best arm which is context *dependent*. This is strongly motivated by the causal setting discussed with real world scenarios in Section 1; (ii) **S2** : This is in accordance with the model in [25]. The contexts and the arms *both can* be chosen by the algorithm and the aim is to compare regret w.r.t the best arm out of *all* KL entries.

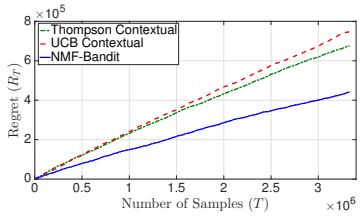
Synthetic Data-Sets : In order to generate the synthetic reward matrix $\mathbf{U}$, the parameter L, K, m are chosen. The $L \times m$ matrix $\mathbf{A}$ is then generated by picking each row uniformly at random from the m-dimensional simplex. The $m \times K$ matrix $\mathbf{W}$ is generated with each entry uniformly generated in the interval $[0, 1]$. We further corrupt 5 % of the entries in each row of $\mathbf{W}$ with completely arbitrary noise while ensuring that they still lie in $[0, 1]$.

In Figure 2a,2b, we compare our algorithm to UCB-1 and Thompson in **S1** under different values of problem dimensions. In Figure 2a, the rewards are uniform with means given by $\mathbf{U}$, while they are Bernoulli in Figure 2b. We observe that UCB-1, Thompson have linear regret as they do not get sufficient concentration for the $L \times K$ mean parameters. However, our algorithm is able to enter the sub-linear regime much faster. We mention the choice of the parameters θ and m' below the corresponding figures. It should be noted that our algorithm performs well even for values of the *explore* parameter θ, which are much lower than prescribed. In Figure 2e the experiments are performed under **S2**. We can see that our algorithm's regret is better compared to the others by a large margin, even though it has not been designed for this setting.

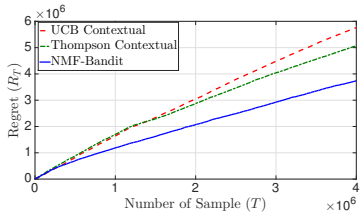
Real World Data-Sets : We use the Movielens 1M [21] and the Book Crossing [41] data-sets for our real world experiments. A subset of dimension 2000×2000 is chosen from the Movielens 1M dataset, such that we have at least 20 ratings in each row and each column. Similarly a subset of 3000×3000 is chosen from the Book Crossing data-set with the same property. Both these partially incomplete rating matrices are then completed using the Python package *fancyimpute* [22] using the default settings. These completed matrices



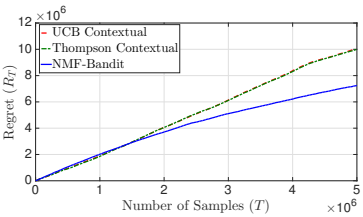
(a) Synthetic data-set with $L = 455$, $K = 210$ and $m = 7$. The rewards are Uniform around the means with a support of length 0.4. Setting : **S1**



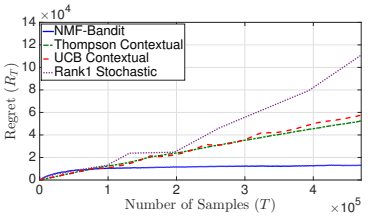
(b) Synthetic data-set with $L = 300$, $K = 145$ and $m = 3$; the rewards are Bernoulli with the given means. Setting : **S1**



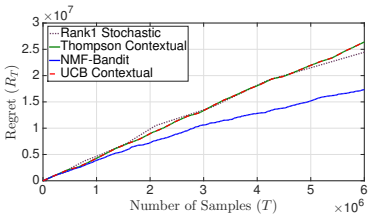
(c) A random subset of Movielens data-set with $L = 1600$, $K = 750$. The rewards are Uniform around the means with a support of length 2. In our algorithm we set $m = 10, m' = 20$ and $\theta = 3$. Setting : **S1**



(d) A random subset of Book Crossing data-set with $L = 3000$, $K = 1450$. The rewards are Uniform around the means with a support of length 2. We set $m = 15, m' = 30$ and $\theta = 3$. Setting: **S1**



(e) Synthetic data-set with $L = 90$, $K = 30$ and $m = 3$. The rewards are Uniform around the means with a support of length 0.4. Setting : **S2**



(f) A random subset of Book Crossing data-set with $L = 1000$, $K = 450$. The rewards are Uniform around the means with a support of length 4. Setting : **S2**

Figure 2: Comparison of contextual versions of UCB-1,Thompson sampling and Rank 1 Stochastic bandits with Algorithm 1 (NMF-Bandit) in **S1** and **S2** on real and synthetic data-sets.

are used in place of the reward matrix $\mathbf{U}$ without any further modifications, and all the algorithms are completely agnostic to the process through which these matrices have been completed. The experiments have been performed in a setting where the rewards observed are uniform around the given means. The support of the uniform distributions are given below each figure.

In Figure 2c and 2d, we compare our algorithm to UCB-1 and Thompson in **S1** on the MovieLens and Book Crossing data-set respectively. As before, our algorithm has superior performance. In Figure 2f, we compare the algorithms on the Book Crossing data-set under **S2**. NMF-Bandit outperforms all the other algorithms, even on the real datasets.

5 Conclusion

In this paper we investigate a causal model of contextual bandits (as shown in Figure 1b) with L observed contexts and K arms, where the observed context influences the reward through a latent confounder. The latent confounder is correlated with the observed context and lies in a lower dimensional space with only m degrees of freedom. We identify that under this causal model, the reward matrix $\mathbf{U}$ naturally factorizes into non-negative factors $\mathbf{A}$ and $\mathbf{W}$.

We propose a novel ϵ-greedy algorithm (NMF-Bandit), which attains a regret guarantee of $O(L\text{poly}(m, \log K) \log T/\Delta^2)$. Our guarantees are under statistical RIP like conditions on the non-negative factors. We also establish a lower bound of $O(Km \log T/\Delta)$ for our problem. To the best of our knowledge, this is the first achievable regret guarantee for online matrix completion with bandit feedback, when rank is greater than one.

This work opens up the prospect of investigating general causal models from a bandit perspective, where the goal is to control the regret of a target variable, when the algorithm can intervene only on some of the variables (*limited interventional capacity*), while other variables (possibly *latent*) can causally influence the reward.

Acknowledgements

This work is partially supported by NSF Grants CNS-1161868, CNS-1343383, CNS-1320175, ARO grant W911NF-16-1-0377 and the US DoT supported D-STOP Tier 1 University Transportation Center.

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] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[4] 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.

[5] 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.

[6] Alexander Barg, Arya Mazumdar, and Rongrong Wang. Restricted isometry property of random subdictionaries. *IEEE Transactions on Information Theory*, 61(8):4440–4450, 2015.

[7] R. Bell, Y. Koren, and C. Volinsky. The bellkor solution to the netflix prize, 2007.

[8] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

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

[10] Stéphane Chrétien and Sébastien Darses. Invertibility of random submatrices via tail-decoupling and a matrix chernoff inequality. *Statistics & Probability Letters*, 82(7):1479–1487, 2012.

[11] Stéphane Chrétien and Zhen Wai Olivier Ho. Small coherence implies the weak null space property. *arXiv preprint arXiv:1606.09193*, 2016.

[12] 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.

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

[14] 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.

[15] 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.

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

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

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

[19] 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.

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

[21] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

[22] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

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

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

[25] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. *arXiv preprint arXiv:1608.03023*, 2016.

[26] E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 592–600, 2012.

[27] 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.

[28] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[29] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *arXiv preprint arXiv:1606.03203*, 2016.

[30] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *The 39th International ACM SIGIR Conference on Information Retrieval (SIGIR)*, 2016.

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

[32] Judea Pearl. *Causality*. Cambridge university press, 2009.

[33] 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.

[34] 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.

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

[36] Joel A Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346(23):1271–1274, 2008.

[37] Joel A Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

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

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

[40] 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.

[41] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.