REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
03-01-2018	Final Report	13-Jul-2016 - 12-Apr-2017

4. TITLE AND SUBTITLE
Final Report: Optimizing Human Input in Social Network Analysis (Topic 10.1.4 - Human Networks)

5a. CONTRACT NUMBER
W911NF-16-1-0377

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas at Austin
101 East 27th Street
Suite 5.300
Austin, TX 78712 -1532

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69165-NS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sanjay Shakkottai
UU	UU	UU	UU		19b. TELEPHONE NUMBER 512-471-5376

RPPR Final Report
as of 23-Jan-2018

Agency Code:

Proposal Number: 69165NSII **Agreement Number: W911NF-16-1-0377**
INVESTIGATOR(S):

Name: Sanjay Shakkottai
Email: shakkott@mail.utexas.edu
Phone Number: 5124715376
Principal: Y

Organization: **University of Texas at Austin**
Address: 101 East 27th Street, Austin, TX 787121532
Country: USA
DUNS Number: 170230239 EIN: 746000203
Report Date: 12-Jul-2017 Date Received: 03-Jan-2018
Final Report for Period Beginning 13-Jul-2016 and Ending 12-Apr-2017
Title: Optimizing Human Input in Social Network Analysis (Topic 10.1.4 - Human Networks)
Begin Performance Period: 13-Jul-2016 **End Performance Period:** 12-Apr-2017
Report Term: 0-Other
Submitted By: Sanjay Shakkottai Email: shakkott@mail.utexas.edu
Phone: (512) 471-5376
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:** 0

Major Goals: The study focused on developing new bandit algorithms for online optimization (e.g. matching tasks to human agents). The attached technical reports provide details on the formulations and results. Specifically, the technical reports focussed on a backlog minimization formulation for matching tasks and agents, as well as a contextual bandit formulation focusing on dimensionality reduction.

Accomplishments: We developed new algorithms for regret minimization, theoretically justified these algorithms' performance, and verified their performance using synthetic and real-world data. We further developed preliminary bandit approaches to using side information, where one arm reveals information on other arms' performance using importance sampling approaches.

The associated papers were published in premier conferences in machine learning (NIPS 2016 and AISTATS 2017).

Training Opportunities: Graduate students were involved in the research.

Results Dissemination: Two papers were published (please see attached technical reports).

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Sanjay Shakkottai
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

RPPR Final Report
as of 23-Jan-2018

Participant Type: Graduate Student (research assistant)
Participant: Rajat Sen
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Regret of Queueing Bandits

Subhashini Krishnasamy
University of Texas at Austin

Rajat Sen
University of Texas at Austin

Ramesh Johari
Stanford University

Sanjay Shakkottai
University of Texas at Austin

Abstract

We consider a variant of the multiarmed bandit problem where jobs *queue* for service, and service rates of different servers may be unknown. We study algorithms that minimize *queue-regret*: the (expected) difference between the queue-lengths obtained by the algorithm, and those obtained by a "genie"-aided matching algorithm that knows exact service rates. A naive view of this problem would suggest that queue-regret should grow logarithmically: since queue-regret cannot be larger than classical regret, results for the standard MAB problem give algorithms that ensure queue-regret increases no more than logarithmically in time. Our paper shows surprisingly more complex behavior. In particular, the naive intuition is correct as long as the bandit algorithm's queues have relatively long regenerative cycles: in this case queue-regret is similar to cumulative regret, and scales (essentially) logarithmically. However, we show that this "early stage" of the queueing bandit eventually gives way to a "late stage", where the optimal queue-regret scaling is $O(1/t)$. We demonstrate an algorithm that (order-wise) achieves this asymptotic queue-regret, and also exhibits close to optimal switching time from the early stage to the late stage.

1 Introduction

Stochastic multi-armed bandits (MAB) have a rich history in sequential decision making [1, 2, 3]. In its simplest form, a collection of K arms are present, each having a binary reward (Bernoulli random variable over $\{0, 1\}$) with an unknown success probability[1] (and different across arms). At each (discrete) time, a single arm is chosen by the bandit algorithm, and a (binary-valued) reward is accrued. The MAB problem is to determine which arm to choose at each time in order to minimize the cumulative expected regret, namely, the cumulative loss of reward when compared to a genie that has knowledge of the arm success probabilities.

In this paper, we consider the variant of this problem motivated by *queueing* applications. Formally, suppose that arms are pulled upon arrivals of *jobs*; each arm is now a *server* that can serve the arriving job. In this model, the stochastic reward described above is equivalent to *service*. In other words, if the arm (server) that is chosen results in positive reward, the job is successfully completed and departs the system. However, this basic model fails to capture an essential feature of service in many settings: in a queueing system, *jobs wait until they complete service*. Such systems are *stateful*: when the chosen arm results in zero reward, the job being served remains in the queue, and over time the model must track the remaining jobs waiting to be served. The difference between the cumulative number of arrivals and departures, or the *queue length*, is the most common measure of the quality of the service strategy being employed.

[1] Here, the success probability of an arm is the probability that the reward equals '1'.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Queueing is employed in modeling a vast range of service systems, including supply and demand in online platforms (e.g., Uber, Lyft, Airbnb, Upwork, etc.); order flow in financial markets (e.g., limit order books); packet flow in communication networks; and supply chains. In all of these systems, queueing is an essential part of the model: e.g., in online platforms, the available supply (e.g. available drivers in Uber or Lyft, or available rentals in Airbnb) queues until it is "served" by arriving demand (ride requests in Uber or Lyft, booking requests in Airbnb). Since MAB models are a natural way to capture learning in this entire range of systems, incorporating queueing behavior into the MAB model is an essential challenge.

This problem clearly has the explore-exploit tradeoff inherent in the standard MAB problem: since the success probabilities across different servers are unknown, there is a tradeoff between learning (*exploring*) the different servers and (*exploiting*) the most promising server from past observations. We refer to this problem as the *queueing bandit*. Since the queue length is simply the difference between the cumulative number arrivals and departures (cumulative actual reward; here reward is 1 if job is served), the natural notion of regret here is to compare the expected queue length under a bandit algorithm with the corresponding one under a genie policy (with identical arrivals) that however always chooses the arm with the highest expected reward.

Queueing System: To capture this trade-off, we consider a discrete-time queueing system with a single queue and K servers. Arrivals to the queue and service offered by the links are according to product Bernoulli distribution and i.i.d. across time slots. Statistical parameters corresponding to the service distributions are considered unknown. In any time slot, the queue can be served by at most one server and the problem is to schedule a server in every time slot. The service is *pre-emptive* and a job returns to the queue if not served. There is at least one server that has a service rate higher than the arrival rate, which ensures that the "genie" policy is stable.

Let $Q(t)$ be the queue length at time t under a given bandit algorithm, and let $Q^*(t)$ be the corresponding queue length under the "genie" policy that always schedules the optimal server (i.e. always plays the arm with the highest mean). We define the *queue-regret* as the difference in expected queue lengths for the two policies. That is, the regret is given by:

$$\Psi(t) := \mathbb{E}\left[Q(t) - Q^*(t)\right]. \tag{1}$$

Here $\Psi(t)$ has the interpretation of the traditional MAB regret with caveat that rewards are accumulated only if there is a job that can benefit from this reward. We refer to $\Psi(t)$ as the *queue-regret*; formally, our goal is to develop bandit algorithms that minimize the queue-regret at a finite time t.

To develop some intuition, we compare this to the standard stochastic MAB problem. For the standard problem, well-known algorithms such as UCB, KL-UCB, and Thompson sampling achieve a cumulative regret of $O((K-1)\log t)$ at time t [4, 5, 6], and this result is essentially tight [7]. In the queueing bandit, we can obtain a simple bound on the queue-regret by noting that it cannot be any higher than the traditional regret (where a reward is accrued at each time whether a job is present or not). This leads to an upper bound of $O((K-1)\log t)$ for the queue regret.

However, this upper bound does not tell the whole story for the queueing bandit: we show that there are two "stages" to the queueing bandit. In the *early* stage, the bandit algorithm is unable to even stabilize the queue – i.e. on average, the queue length increases over time and is continuously backlogged; therefore the queue-regret grows with time, similar to the cumulative regret. Once the algorithm is able to stabilize the queue—the *late* stage—then a dramatic shift occurs in the behavior of the queue regret. A stochastically stable queue goes through **regenerative cycles** – a random cyclical behavior where queues build-up over time, then empty, and the cycle repeats. The associated recurring"zero-queue-length" epochs means that sample-path queue-regret essentially "resets" at (stochastically) regular intervals; i.e., the sample-path queue-regret becomes non-positive at these time instants. Thus the queue-regret should fall over time, as the algorithm learns.

Our main results provide lower bounds on queue-regret for both the early and late stages, as well as algorithms that essentially match these lower bounds. We first describe the late stage, and then describe the early stage for a heavily loaded system.

1. The late stage. We first consider what happens to the queue regret as $t \to \infty$. As noted above, a reasonable intuition for this regime comes from considering a standard bandit algorithm, but where the sample-path queue-regret "resets" at time points of regeneration.[2] In this case, the queue-regret is

[2]This is inexact since the optimal queueing system and bandit queueing system may not regenerate at the same time point; but the intuition holds.

approximately a (discrete) *derivative* of the cumulative regret. Since the optimal cumulative regret scales like $\log t$, asymptotically the optimal queue-regret should scale like $1/t$. Indeed, we show that the queue-regret for α-consistent policies is at least C/t infinitely often, where C is a constant independent of t. Further, we introduce an algorithm called Q-ThS for the queueing bandit (a variant of Thompson sampling with explicit structured exploration), and show an asymptotic regret upper bound of $O\left(\text{poly}(\log t)/t\right)$ for Q-ThS, thus matching the lower bound up to poly-logarithmic factors in t. Q-ThS exploits *structured exploration*: we exploit the fact that the queue regenerates regularly to explore more systematically and aggressively.

2. The early stage. The preceding discussion might suggest that an algorithm that explores aggressively would dominate any algorithm that balances exploration and exploitation. However, an overly aggressive exploration policy will preclude the queueing system from ever stabilizing, which is *necessary* to induce the regenerative cycles that lead the system to the late stage. To even enter the late stage, therefore, we need an algorithm that exploits enough to actually stabilize the queue (i.e. choose good arms sufficiently often so that the mean service rate exceeds the expected arrival rate).

We refer to the early stage of the system, as noted above, as the period before the algorithm has learned to stabilize the queues. For a *heavily loaded system, where the arrival rate approaches the service rate of the optimal server,* we show a lower bound of $\Omega(\log t/\log\log t)$ on the queue-regret in the early stage. Thus up to a $\log\log t$ factor, the early stage regret behaves similarly to the cumulative regret (which scales like $\log t$). The heavily loaded regime is a natural asymptotic regime in which to study queueing systems, and has been extensively employed in the literature; see, e.g., [8, 9] for surveys.

Perhaps more importantly, our analysis shows that the time to switch from the early stage to the late stage scales at least as $t = \Omega(K/\epsilon)$, where ϵ is the gap between the arrival rate and the service rate of the optimal server; thus $\epsilon \to 0$ in the heavy-load setting. In particular, we show that the early stage lower bound of $\Omega(\log t/\log\log t)$ is valid up to $t = O(K/\epsilon)$; on the other hand, we also show that, in the heavy-load limit, depending on the relative scaling between K and ϵ, the regret of Q-ThS scales like $O\left(\text{poly}(\log t)/\epsilon^2 t\right)$ for times that are arbitrarily close to $\Omega(K/\epsilon)$. In other words, Q-ThS is nearly optimal in the time it takes to "switch" from the early stage to the late stage.

Our results constitute the first insight into the behavior of regret in this queueing setting; as emphasized, it is quite different than that seen for minimization of cumulative regret in the standard MAB problem. The preceding discussion highlights why minimization of queue-regret presents a subtle learning problem. On one hand, if the queue has been stabilized, the presence of regenerative cycles allows us to establish that queue regret must eventually decay to zero at rate $1/t$ under an optimal algorithm (the late stage). On the other hand, to actually have regenerative cycles in the first place, a learning algorithm needs to exploit enough to actually stabilize the queue (the early stage). Our analysis not only characterizes regret in both regimes, but also essentially exactly characterizes the transition point between the two regimes. In this way the queueing bandit is a remarkable new example of the tradeoff between exploration and exploitation.

2 Related work

MAB algorithms. Stochastic MAB models have been widely used in the past as a paradigm for various sequential decision making problems in industrial manufacturing, communication networks, clinical trials, online advertising and webpage optimization, and other domains requiring resource allocation and scheduling; see, e.g., [1, 2, 3]. The MAB problem has been studied in two variants, based on different notions of optimality. One considers mean accumulated loss of rewards, often called *regret*, as compared to a genie policy that always chooses the best arm. Most effort in this direction is focused on getting the best regret bounds possible at any *finite time* in addition to designing computationally feasible algorithms [3]. The other line of research models the bandit problem as a Markov decision process (MDP), with the goal of optimizing *infinite horizon* discounted or average reward. The aim is to characterize the structure of the optimal policy [2]. Since these policies deal with optimality with respect to infinite horizon costs, unlike the former body of research, they give steady-state and not finite-time guarantees. Our work uses the regret minimization framework to study the queueing bandit problem.

Bandits for queues. There is body of literature on the application of bandit models to queueing and scheduling systems [2, 10, 11, 12, 13, 14, 15, 16]. These queueing studies focus on infinite-horizon

costs (i.e., statistically steady-state behavior, where the focus typically is on conditions for optimality of index policies); further, the models do not typically consider user-dependent server statistics. Our focus here is different: algorithms and analysis to optimize finite time regret.

3 Problem Setting

We consider a discrete-time queueing system with a single queue and K servers. The servers are indexed by $k = 1, \ldots, K$. Arrivals to the queue and service offered by the links are according to product Bernoulli distribution and i.i.d. across time slots. The mean arrival rate is given by λ and the mean service rates by the vector $\boldsymbol{\mu} = [\mu_k]_{k\in[K]}$, with $\lambda < \max_{k\in[K]} \mu_k$. In any time slot, the queue can be served by at most one server and the problem is to schedule a server in every time slot. The scheduling decision at any time t is based on past observations corresponding to the services obtained from the scheduled servers until time $t-1$. Statistical parameters corresponding to the service distributions are considered unknown. The queueing system evolution can be described as follows. Let $\kappa(t)$ denote the server that is scheduled at time t. Also, let $R_k(t) \in \{0, 1\}$ be the service offered by server k and $S(t)$ denote the service offered by server $\kappa(t)$ at time t, i.e., $S(t) = R_{\kappa(t)}(t)$. If $A(t)$ is the number of arrivals at time t, then the queue-length at time t is given by: $Q(t) = (Q(t-1) + A(t) - S(t))^+$.

Our goal in this paper is to focus attention on how queueing behavior impacts regret minimization in bandit algorithms. We evaluate the performance of scheduling policies against the policy that schedules the (unique) optimal server in every time slot, i.e., the server $k^* := \arg\max_{k\in[K]} \mu_k$ with the maximum mean rate $\mu^* := \max_{k\in[K]} \mu_k$. Let $Q(t)$ be the queue-length vector at time t under our specified algorithm, and let $Q^*(t)$ be the corresponding vector under the optimal policy. We define *regret* as the difference in mean queue-lengths for the two policies. That is, the regret is given by: $\Psi(t) := \mathbb{E}\left[Q(t) - Q^*(t)\right]$. We use the terms *queue-regret* or simply *regret* to refer to $\Psi(t)$.

Throughout, when we evaluate queue-regret, we do so under the assumption that the queueing system starts in the steady state distribution of the system induced by the optimal policy, as follows.

Assumption 1 (Initial State)**.** *Both $Q(0)$ and $Q^*(0)$ have the same initial state distribution, and this is chosen to be the stationary distribution of $Q^*(t)$; this distribution is denoted $\pi_{(\lambda,\mu^*)}$.*

4 The Late Stage

We analyze the performance of a scheduling algorithm with respect to queue-regret as a function of time and system parameters like: *(a)* the load on the system $\epsilon := (\mu^* - \lambda)$, and *(b)* the minimum difference between the rates of the best and the next best servers $\Delta := \mu^* - \max_{k\neq k^*} \mu_k$.

As a preview of the theoretical results, Figure 1 shows the evolution of queue-regret with time in a system with 5 servers under a scheduling policy inspired by Thompson Sampling. Exact details of the scheduling algorithm can be found in Section 4.2. It is observed that the regret goes through a phase transition. In the initial stage, when the algorithm has not estimated the service rates well enough to stabilize the queue, the regret grows poly-logarithmically similar to the classical MAB setting. After a critical point when the algorithm has learned the system parameters well enough to stabilize the queue, the queue-length goes through regenerative cycles as the queue become empty. In other-words, instead of the queue length being continuously backlogged, the queuing system has a stochastic cyclical behavior where the queue builds up, becomes empty, and this cycle recurs. Thus at the beginning of every regenerative cycle, there is no accumulation of past errors and the sample-path queue-regret is at most zero. As the algorithm estimates the parameters better with time, the length of the regenerative cycles decreases and the queue-regret decays to zero.

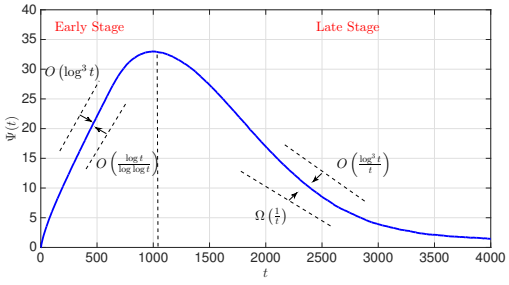


Figure 1: Queue-regret $\Psi(t)$ under Q-ThS in a system with $K = 5$, $\epsilon = 0.1$ and $\Delta = 0.17$

Notation: For the results in Section 4, the notation $f(t) = O\left(g(K, \epsilon, t)\right)$ for all $t \in h(K, \epsilon)$ (here, $h(K, \epsilon)$ is an interval that depends on K, ϵ) implies that there exist constants C and t_0 independent of K and ϵ such that $f(t) \leq Cg(K, \epsilon, t)$ for all $t \in (t_0, \infty) \cap h(K, \epsilon)$.

4.1 An Asymptotic Lower Bound

We establish an asymptotic lower bound on regret for the class of α-consistent policies; this class for the queueing bandit is a generalization of the α-consistent class used in the literature for the traditional stochastic MAB problem [7, 17, 18]. The precise definition is given below ($\mathbb{1}\{\cdot\}$ below is the indicator function).

Definition 1. *A scheduling policy is said to be α-consistent (for some $\alpha \in (0, 1)$) if given any problem instance, specified by $(\lambda, \boldsymbol{\mu})$, $\mathbb{E}\left[\sum_{s=1}^{t} \mathbb{1}\{\kappa(s) = k\}\right] = O(t^{\alpha})$ for all $k \neq k^*$.*

Theorem 1 below gives an asymptotic lower bound on the average queue-regret and per-queue regret for an arbitrary α-consistent policy.

Theorem 1. *For any problem instance $(\lambda, \boldsymbol{\mu})$ and any α-consistent policy, the regret $\Psi(t)$ satisfies*

$$\Psi(t) \geq \left(\frac{\lambda}{4} D(\boldsymbol{\mu})(1-\alpha)(K-1)\right) \frac{1}{t}$$

for infinitely many t, where

$$D(\boldsymbol{\mu}) = \frac{\Delta}{\mathrm{KL}\left(\mu_{min}, \frac{\mu^*+1}{2}\right)}. \tag{2}$$

Outline for theorem 1. The proof of the lower bound consists of three main steps. First, in lemma 21, we show that the regret at any time-slot is lower bounded by the probability of a sub-optimal schedule in that time-slot (up to a constant factor that is dependent on the problem instance). The key idea in this lemma is to show the equivalence of any two systems with the same marginal service distributions under bandit feedback. This is achieved through a carefully constructed coupling argument that maps the original system with independent service across links to another system with service process that is dependent across links but with the same marginal distribution.

As a second step, the lower bound on the regret in terms of the probability of a sub-optimal schedule enables us to obtain a lower bound on the cumulative queue-regret in terms of the number of sub-optimal schedules. We then use a lower bound on the number of sub-optimal schedules for α-consistent policies (lemma 19 and corollary 20) to obtain a lower bound on the cumulative regret. In the final step, we use the lower bound on the cumulative queue-regret to obtain an *infinitely often* lower bound on the queue-regret. □

4.2 Achieving the Asymptotic Bound

We next focus on algorithms that can (up to a poly log factor) achieve a scaling of $O\left(1/t\right)$. A key challenge in showing this is that we will need high probability bounds on the number of times the correct arm is scheduled, and these bounds to hold over the late-stage regenerative cycles of the queue. Recall that these regenerative cycles are random time intervals with $\Theta(1)$ expected length for the optimal policy, and whose lengths are correlated with the bandit algorithm decisions (the queue length evolution is dependent on the past history of bandit arm schedules). To address this, we propose a slightly modified version of the Thompson Sampling algorithm. The algorithm, which we call Q-ThS, has an explicit structured exploration component similar to ϵ-greedy algorithms. This structured exploration provides sufficiently good estimates for all arms (including sub-optimal ones) in the late stage.

We describe the algorithm we employ in detail. Let $T_k(t)$ be the number of times server k is assigned in the first t time-slots and $\hat{\boldsymbol{\mu}}(t)$ be the empirical mean of service rates at time-slot t from past observations (until $t-1$). At time-slot t, Q-ThS decides to *explore* with probability $\min\{1, 3K \log^2 t/t\}$, otherwise it *exploits*. When exploring, it chooses a server uniformly at random. The chosen exploration rate ensures that we are able to obtain concentration results for the number

of times any link is sampled.[3] When exploiting, for each $k \in [K]$, we pick a sample $\hat{\theta}_k(t)$ of distribution $\text{Beta}\left(\hat{\mu}_k(t)T_k(t-1)+1, (1-\hat{\mu}_k(t))\,T_k(t-1)+1\right)$, and schedule the arm with the largest sample (the standard Thompson sampling for Bernoulli arms [19]). Details of the algorithm are given in Algorithm 1 in the Appendix.

We now show that, for a given problem instance $(\lambda, \boldsymbol{\mu})$ (and therefore fixed ϵ), the regret under Q-ThS scales as $O\left(\text{poly}(\log t)/t\right)$. We state the most general form of the asymptotic upper bound in theorem 2. A slightly weaker version of the result is given in corollary 3. This corollary is useful to understand the dependence of the upper bound on the load ϵ and the number of servers K.

Notation : For the following results, the notation $f(t) = O\left(g(K,\epsilon,t)\right)$ for all $t \in h(K,\epsilon)$ (here, $h(K,\epsilon)$ is an interval that depends on K, ϵ) implies that there exist constants C and t_0 independent of K and ϵ such that $f(t) \leq Cg(K,\epsilon,t)$ for all $t \in (t_0, \infty) \cap h(K,\epsilon)$.

Theorem 2. *Consider any problem instance $(\lambda, \boldsymbol{\mu})$. Let $w(t) = \exp\left(\left(\frac{2\log t}{\Delta}\right)^{2/3}\right)$, $v'(t) = \frac{6K}{\epsilon}w(t)$ and $v(t) = \frac{24}{\epsilon^2}\log t + \frac{60K}{\epsilon}\frac{v'(t)\log^2 t}{t}$. Then, under Q-ThS the regret $\Psi(t)$, satisfies*

$$\Psi(t) = O\left(\frac{Kv(t)\log^2 t}{t}\right)$$

for all t such that $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon}$, $t \geq \exp\left(6/\Delta^2\right)$ and $v(t) + v'(t) \leq t/2$.

Corollary 3. *Let $w(t)$ be as defined in Theorem 2. Then,*

$$\Psi(t) = O\left(K\frac{\log^3 t}{\epsilon^2 t}\right)$$

for all t such that $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon}$, $\frac{t}{w(t)} \geq \max\left\{\frac{24K}{\epsilon}, 15K^2\log t\right\}$, $t \geq \exp\left(6/\Delta^2\right)$ and $\frac{t}{\log t} \geq \frac{198}{\epsilon^2}$.

Outline for Theorem 2. As mentioned earlier, the central idea in the proof is that the sample-path queue-regret is at most zero at the beginning of regenerative cycles, i.e., instants at which the queue becomes empty. The proof consists of two main parts – one which gives a high probability result on the number of sub-optimal schedules in the exploit phase in the late stage, and the other which shows that at any time, the beginning of the current regenerative cycle is not very far in time.

The former part is proved in lemma 9, where we make use of the structured exploration component of Q-ThS to show that all the links, including the sub-optimal ones, are sampled a sufficiently large number of times to give a good estimate of the link rates. This in turn ensures that the algorithm schedules the correct link in the exploit phase in the late stages with high probability.

For the latter part, we prove a high probability bound on the last time instant when the queue was zero (which is the beginning of the current regenerative cycle) in lemma 15. Here, we make use of a recursive argument to obtain a tight bound. More specifically, we first use a coarse high probability upper bound on the queue-length (lemma 11) to get a first cut bound on the beginning of the regenerative cycle (lemma 12). This bound on the regenerative cycle-length is then recursively used to obtain tighter bounds on the queue-length, and in turn, the start of the current regenerative cycle (lemmas 14 and 15 respectively).

The proof of the theorem proceeds by combining the two parts above to show that the main contribution to the queue-regret comes from the structured exploration component in the current regenerative cycle, which gives the stated result. □

5 The Early Stage in the Heavily Loaded Regime

In order to study the performance of α-consistent policies in the early stage, we consider the *heavily loaded* system, where the arrival rate λ is close to the optimal service rate μ^*, i.e., $\epsilon = \mu^* - \lambda \to 0$. This is a well studied asymptotic in which to study queueing systems, as this regime leads to

[3] The exploration rate could scale like $\log t/t$ if we knew Δ in advance; however, without this knowledge, additional exploration is needed.

fundamental insight into the structure of queueing systems. See, e.g., [8, 9] for extensive surveys. Analyzing queue-regret in the early stage in the heavily loaded regime has the effect that the the optimal server is the only one that stabilizes the queue. As a result, in the heavily loaded regime, effective learning and scheduling of the optimal server play a crucial role in determining the transition point from the early stage to the late stage. For this reason the heavily loaded regime reveals the behavior of regret in the early stage.

Notation: For all the results in this section, the notation $f(t) = O\left(g(K, \epsilon, t)\right)$ for all $t \in h(K, \epsilon)$ ($h(K, \epsilon)$ is an interval that depends on K, ϵ) implies that there exist numbers C and ϵ_0 that depend on Δ such that for all $\epsilon \geq \epsilon_0$, $f(t) \leq Cg(K, \epsilon, t)$ for all $t \in h(K, \epsilon)$.

Theorem 4 gives a lower bound on the regret in the heavily loaded regime, roughly in the time interval $\left(K^{1/(1-\alpha)}, O\left(K/\epsilon\right)\right)$ for any α-consistent policy.

Theorem 4. *Given any problem instance $(\lambda, \boldsymbol{\mu})$, and for any α-consistent policy and $\gamma > \frac{1}{1-\alpha}$, the regret $\Psi(t)$ satisfies*

$$\Psi(t) \geq \frac{D(\boldsymbol{\mu})}{2}(K-1)\frac{\log t}{\log\log t}$$

for $t \in \left[\max\{C_1 K^{\gamma}, \tau\}, (K-1)\frac{D(\boldsymbol{\mu})}{2\epsilon}\right]$ where $D(\boldsymbol{\mu})$ is given by equation 2, and τ and C_1 are constants that depend on α, γ and the policy.

Outline for Theorem 4. The crucial idea in the proof is to show a lower bound on the queue-regret in terms of the number of sub-optimal schedules (Lemma 22). As in Theorem 1, we then use a lower bound on the number of sub-optimal schedules for α-consistent policies (given by Corollary 20) to obtain a lower bound on the queue-regret. □

Theorem 4 shows that, for any α-consistent policy, it takes at least $\Omega\left(K/\epsilon\right)$ time for the queue-regret to transition from the early stage to the late stage. In this region, the scaling $O(\log t/\log\log t)$ reflects the fact that queue-regret is dominated by the cumulative regret growing like $O(\log t)$. A reasonable question then arises: after time $\Omega\left(K/\epsilon\right)$, should we expect the regret to transition into the late stage regime analyzed in the preceding section?

We answer this question by studying when Q-ThS achieves its late-stage regret scaling of $O\left(\text{poly}(\log t)/\epsilon^2 t\right)$ scaling; as we will see, in an appropriate sense, Q-ThS is close to optimal in its transition from early stage to late stage, when compared to the bound discovered in Theorem 4. Formally, we have Corollary 5, which is an analog to Corollary 3 under the heavily loaded regime.

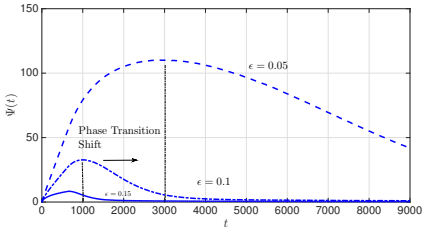
Corollary 5. *For any problem instance $(\lambda, \boldsymbol{\mu})$, any $\gamma \in (0, 1)$ and $\delta \in (0, \min(\gamma, 1-\gamma))$, the regret under Q-ThS satisfies*

$$\Psi(t) = O\left(\frac{K\log^3 t}{\epsilon^2 t}\right)$$

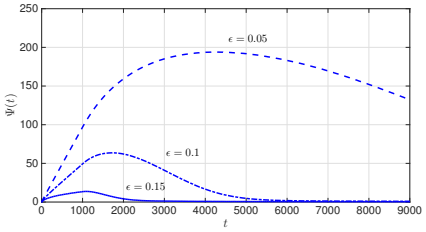
$\forall t \geq C_2 \max\left\{\left(\frac{1}{\epsilon}\right)^{\frac{1}{\gamma-\delta}}, \left(\frac{K}{\epsilon}\right)^{\frac{1}{1-\gamma}}, (K^2)^{\frac{1}{1-\gamma-\delta}}, \left(\frac{1}{\epsilon^2}\right)^{\frac{1}{1-\delta}}\right\}$, where C_2 is a constant independent of ϵ (but depends on Δ, γ and δ).

By combining the result in Corollary 5 with Theorem 4, we can infer that in the heavily loaded regime, the time taken by Q-ThS to achieve $O\left(\text{poly}(\log t)/\epsilon^2 t\right)$ scaling is, in some sense, order-wise close to the optimal in the α-consistent class. Specifically, for any $\beta \in (0, 1)$, there exists a scaling of K with ϵ such that the queue-regret under Q-ThS scales as $O\left(\text{poly}(\log t)/\epsilon^2 t\right)$ for all $t > (K/\epsilon)^{\beta}$ while the regret under any α-consistent policy scales as $\Omega\left(K\log t/\log\log t\right)$ for $t < K/\epsilon$.

We conclude by noting that while the transition point from the early stage to the late stage for Q-ThS is near optimal in the heavily loaded regime, it does not yield optimal regret performance in the early stage in general. In particular, recall that at any time t, the structured exploration component in Q-ThS is invoked with probability $3K\log^2 t/t$. As a result, we see that, in the early stage, queue-regret under Q-ThS could be a $\log^2 t$-factor worse than the $\Omega\left(\log t/\log\log t\right)$ lower bound shown in Theorem 4 for the α-consistent class. This intuition can be formalized: it is straightforward to show an upper bound of $2K\log^3 t$ for any $t > \max\{C_3, U\}$, where C_3 is a constant that depends on Δ but is independent of K and ϵ; we omit the details.



(a) Queue-Regret under Q-ThS for a system with 5 servers with $\epsilon \in \{0.05, 0.1, 0.15\}$



(b) Queue-Regret under Q-ThS for a a system with 7 servers with $\epsilon \in \{0.05, 0.1, 0.15\}$

Figure 2: Variation of Queue-regret $\Psi(t)$ with K and ϵ under Q-Ths. The phase-transition point shifts towards the right as ϵ decreases. The efficiency of learning decreases with increase in the size of the system.

6 Simulation Results

In this section we present simulation results of various queueing bandit systems with K servers. These results corroborate our theoretical analysis in Sections 4 and 5. In particular a phase transition from unstable to stable behavior can be observed in all our simulations, as predicted by our analysis. In the remainder of the section we demonstrate the performance of Algorithm 1 under variations of system parameters like the traffic (ϵ), the gap between the optimal and the suboptimal servers (Δ), and the size of the system (K). We also compare the performance of our algorithm with versions of UCB-1 [4] and Thompson Sampling [19] without structured exploration (Figure 3 in the appendix).

Variation with ϵ and K. In Figure 2 we see the evolution of $\Psi(t)$ in systems of size 5 and 7 . It can be observed that the regret decays faster in the smaller system, which is predicted by Theorem 2 in the late stage and Corollary 5 in the early stage. The performance of the system under different traffic settings can be observed in Figure 2. It is evident that the regret of the queueing system grows with decreasing ϵ. This is in agreement with our analytical results (Corollaries 3 and 5). In Figure 2 we can observe that the time at which the phase transition occurs shifts towards the right with decreasing ϵ which is predicted by Corollaries 3 and 5.

7 Discussion and Conclusion

This paper provides the first regret analysis of the queueing bandit problem, including a characterization of regret in both early and late stages, together with analysis of the switching time; and an algorithm (Q-ThS) that is asymptotically optimal (to within poly-logarithmic factors) and also essentially exhibits the correct switching behavior between early and late stages. There remain substantial open directions for future work.

First, is there a single algorithm that gives optimal performance in *both* early and late stages, as well as the optimal switching time between early and late stages? The price paid for structured exploration by Q-ThS is an inflation of regret in the early stage. An important open question is to find a single, adaptive algorithm that gives good performance over all time. As we note in the appendix, classic (unstructured) Thompson sampling is an intriguing candidate from this perspective.

Second the most significant technical hurdle in finding a single optimal algorithm is the difficulty of establishing concentration results for the number of suboptimal arm pulls within a regenerative cycle whose length is dependent on the bandit strategy. Such concentration results would be needed in two different limits: first, as the start time of the regenerative cycle approaches infinity (for the asymptotic analysis of late stage regret); and second, as the load of the system increases (for the analysis of early stage regret in the heavily loaded regime). Any progress on the open directions described above would likely require substantial progress on these technical questions as well.

Acknowledgement: This work is partially supported by NSF Grants CNS-1161868, CNS-1343383, CNS-1320175, ARO grants W911NF-16-1-0377, W911NF-15-1-0227, W911NF-14-1-0387 and the US DoT supported D-STOP Tier 1 University Transportation Center.

References

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

[2] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. Springer, 2008, pp. 121–151.

[3] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[5] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," *arXiv preprint arXiv:1102.2490*, 2011.

[6] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," *arXiv preprint arXiv:1111.1797*, 2011.

[7] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[8] W. Whitt, "Heavy traffic limit theorems for queues: a survey," in *Mathematical Methods in Queueing Theory*. Springer, 1974, pp. 307–350.

[9] H. Kushner, *Heavy traffic analysis of controlled queueing and communication networks*. Springer Science & Business Media, 2013, vol. 47.

[10] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *Top*, vol. 15, no. 2, pp. 161–198, 2007.

[11] P. Jacko, "Restless bandits approach to the job scheduling problem and its extensions," *Modern trends in controlled stochastic processes: theory and applications*, pp. 248–267, 2010.

[12] D. Cox and W. Smith, "Queues," *Wiley*, 1961.

[13] C. Buyukkoc, P. Varaiya, and J. Walrand, "The cμ rule revisited," *Advances in applied probability*, vol. 17, no. 1, pp. 237–238, 1985.

[14] J. A. Van Mieghem, "Dynamic scheduling with convex delay costs: The generalized c mu rule," *The Annals of Applied Probability*, pp. 809–833, 1995.

[15] J. Niño-Mora, "Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue," *Queueing Systems*, vol. 54, no. 4, pp. 281–312, 2006.

[16] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Probability in the Engineering and Informational Sciences*, vol. 14, pp. 259–297, 2000.

[17] A. Salomon, J.-Y. Audiber, and I. El Alaoui, "Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 187–207, 2013.

[18] R. Combes, C. Jiang, and R. Srikant, "Bandits with budgets: Regret lower bounds and optimal algorithms," in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. ACM, 2015, pp. 245–257.

[19] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, pp. 285–294, 1933.

[20] S. Bubeck, V. Perchet, and P. Rigollet, "Bounded regret in stochastic multi-armed bandits," *arXiv preprint arXiv:1302.1611*, 2013.

[21] V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg, "Batched bandit problems," *arXiv preprint arXiv:1505.00369*, 2015.

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

[23] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in neural information processing systems*, 2011, pp. 2249–2257.

[24] S. L. Scott, "A modern bayesian look at the multi-armed bandit," *Appl. Stoch. Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.

[25] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Algorithmic Learning Theory*. Springer, 2012, pp. 199–213.

[26] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

Appendix

Algorithm 1 Q-ThS

At time t,
Let $\mathsf{E}(t)$ be an independent Bernoulli sample of mean $\min\{1, 3K\frac{\log^2 t}{t}\}$.
if $\mathsf{E}(t) = 1$ **then**
Explore:
Schedule a server uniformly at random.
else
Exploit:
For each $k \in [K]$, pick a sample $\hat{\theta}_k(t)$ of distribution,

$$\hat{\theta}_k(t) \sim \text{Beta}\left(\hat{\mu}_k(t)T_k(t-1)+1, (1-\hat{\mu}_k(t))\,T_k(t-1)+1\right).$$

Schedule a server

$$\kappa(t) \in \arg\max_{k\in[K]} \hat{\theta}_k(t).$$

end if

We present our theoretical results in a more general setting where there are U queues and K servers, such that $1 \leq U \leq K$. All the results in the body of the paper become a special case of this setting when $U = 1$. The queues and servers are indexed by $u = 1, \ldots, U$ and $k = 1, \ldots, K$ respectively. Arrivals to queues and service offered by the links are according to product Bernoulli distribution and i.i.d. across time slots. The mean arrival rates are given by the vector $\boldsymbol{\lambda} = (\lambda_u)_{u\in[U]}$ and the mean service rates by the matrix $\boldsymbol{\mu} = [\mu_{uk}]_{u\in[U],k\in[K]}$.

In any time slot, each server can serve at most one queue and each queue can be served by at most one server. The problem is to schedule, in every time slot, a matching in the complete bipartite graph between queues and servers. The scheduling decision at any time t is based on past observations corresponding to the services obtained for the scheduled matchings until time $t-1$. Statistical parameters corresponding to the service distributions are considered unknown. The relevant notation for this system has been provided in Table 1.

Table 1: General Notation

Symbol	Description
λ_u	Expected rate of arrival to queue u
λ_{min}	Minimum arrival rate across all queues
$A_u(t)$	Arrival at time t to queue u
μ_{uk}	Expected service rate of server k for queue u
$R_{uk}(t)$	Service rate between server k queue u at time t
k_u^*	Best server for queue u
μ_u^*	Expected rate of best server for queue u
μ_{max}	Maximum service rate across all links
μ_{min}	Minimum service rate across all links
Δ	Minimum (among all queues) difference between the best and second best servers
$\kappa_u(t)$	server assigned to queue u at time t
$S_u(t)$	Potential service provided by server assigned to queue u at time t
$Q_u(t)$	queue-length of queue u at time t
$Q_u^*(t)$	queue-length of queue u at time t for the optimal strategy
$\Psi_u(t)$	Regret for queue u at time t

The queueing system evolution can be described as follows. Let $\kappa_u(t)$ denote the server that is assigned to queue u at time t. Therefore, the vector $\boldsymbol{\kappa}(t) = (\kappa_u(t)_{u\in[U]})$ gives the matching scheduled at time t. Let $R_{uk}(t)$ be the service offered to queue u by server k and $S_u(t)$ denote the service offered to queue u by server $\kappa_u(t)$ at time t. If $\mathbf{A}(t)$ is the (binary) arrival vector at time t, then the queue-length vector at time t is given by:

$$\mathbf{Q}(t) = (\mathbf{Q}(t-1) + \mathbf{A}(t) - \mathbf{S}(t))^+ .$$

Regret Against a Unique Optimal Matching

Our goal in this paper is to focus attention on how queueing behavior impacts regret minimization in bandit algorithms. To emphasize this point, we consider a somewhat simplified switch scheduling system. In particular, we assume for every queue, there is a unique optimal server with the maximum expected service rate for that queue. Further, we assume that the optimal queue-server pairs form a matching in the complete bipartite graph between queues and servers, that we call the *optimal matching*; and that this optimal matching stabilizes every queue.

Formally, make the following definitions:

$$\mu_u^* := \max_{k\in[K]} \mu_{uk}, \quad u \in [U]; \tag{3}$$

$$k_u^* := \arg\max_{k\in[K]} \mu_{uk}, \quad u \in [U]; \tag{4}$$

$$\epsilon_u := \mu_u^* - \lambda_u, \quad u \in [U]; \tag{5}$$

$$\Delta_{uk} := \mu_u^* - \mu_{uk}, \quad u \in [U], k \in [K]; \tag{6}$$

$$\Delta := \min_{u\in[U],k\notin k_u^*} \Delta_{uk}; \tag{7}$$

$$\mu_{min} := \min_{u\in[U],k\in[K]} \mu_{uk}; \tag{8}$$

$$\mu_{max} := \max_{u\in[U],k\in[K]} \mu_{uk}; \tag{9}$$

$$\lambda_{min} := \min_{u\in[U]} \lambda_u. \tag{10}$$

The following assumptions will be in force throughout the paper.

Assumption 2 (Optimal Matching)**.** *There is a* unique optimal matching*, i.e.:*

1. *There is a unique optimal server for each queue: k_u^* is a singleton, i.e., $\Delta_{uk} > 0$ for $k \neq k_u^*$, for all u,*
2. *The optimal queue-server pairs for a matching: For any $u' \neq u$, $k_u^* \neq k_{u'}^*$.*

Assumption 3 (Stability)**.** *The optimal matching stabilizes every queue, i.e., the arrival rates lie within the stability region: $\epsilon_u > 0$ for all $u \in [U]$.*

The assumption of a unique optimal matching essentially means that the queues and servers are solving a pure coordination problem; for example, in the crowdsourcing example described in the introduction, this would correspond to the presence of a unique worker best suited to each type of job. Note that the setting described in Section 3 is equivalent to the unique optimal matching case when $U = 1$. We now describe an algorithm for the unique best match setting which is a more general version of Algorithm 1.

The notation specific to Algorithm 2 has been provided in Table 2.

8 Proofs

We provide details of the proofs for Theorem 2 in Section 8.1 and for Theorems 16 and 17 in Section 8.2. In each section, we state and prove a few intermediate lemmas that are useful in proving the theorems.

Algorithm 2 Q-ThS(match)

At time t,

Let $\mathsf{E}(t)$ be an independent Bernoulli sample of mean $\min\{1, 3K\frac{\log^2 t}{t}\}$.

if $\mathsf{E}(t) = 1$ **then**

Explore:

Schedule a matching from $\mathcal{E}$ uniformly at random.

else

Exploit:

For each $k \in [K], u \in [U]$, pick a sample $\hat{\theta}_{uk}(t)$ of distribution,

$$\hat{\theta}_{uk}(t) \sim \text{Beta}\left(\hat{\mu}_{uk}(t)T_{uk}(t-1)+1, (1-\hat{\mu}_{uk}(t))\,T_{uk}(t-1)+1\right).$$

Compute for all $u \in [U]$

$$\hat{k}_u(t) := \arg\max_{k\in[K]} \hat{\theta}_{uk}(t)$$

Schedule a matching $\boldsymbol{\kappa}(t)$ such that

$$\boldsymbol{\kappa}(t) \in \arg\min_{\boldsymbol{\kappa}\in\mathcal{M}} \sum_{u\in[U]} \mathbb{1}\left\{\kappa_u \neq \hat{k}_u(t)\right\},$$

i.e., $\boldsymbol{\kappa}(t)$ *is the projection of* $\hat{\boldsymbol{k}}(t)$ *onto the space of all matchings* $\mathcal{M}$ *with Hamming distance as metric.*

end if

Table 2: Notation specific to Algorithm 2

Symbol	Description
$\mathsf{E}(t)$	Indicates if the algorithm schedules a matching through *Explore*
$\mathsf{E}_{uk}(t)$	Indicates if Server k is assigned to Queue u at time t through *Explore*
$\mathsf{I}_{uk}(t)$	Indicates if Server k is assigned to Queue u at time t through *Exploit*
$T_{uk}(t)$	Number of time slots Server k is assigned to Queue u in time $[1, t]$
$\hat{\boldsymbol{\mu}}(t)$	Empirical mean of service rates at time t from past observations (until $t-1$)
$\boldsymbol{\kappa}(t)$	Matching scheduled in time-slot t

8.1 Regret Upper Bound for Q-ThS(match)

Theorem 2 is a special case ($U = 1$) of Theorem 6 stated below,

Theorem 6. *Consider any problem instance* $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ *which has a single best matching. For any* $u \in [U]$, *let* $w(t) = \exp\left(\left(\frac{2\log t}{\Delta}\right)^{2/3}\right)$, $v'_u(t) = \frac{6K}{\epsilon_u} w(t)$, $t \geq \exp\left(6/\Delta^2\right)$ *and* $v_u(t) = \frac{24}{\epsilon_u^2}\log t + \frac{60K}{\epsilon_u}\frac{v'_u(t)\log^2 t}{t}$. *Then, under Q-ThS(match) the regret for queue* u, $\Psi_u(t)$, *satisfies*

$$\Psi_u(t) = O\left(\frac{K v_u(t)\log^2 t}{t}\right)$$

for all t *such that* $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$, $t \geq \exp\left(6/\Delta^2\right)$ *and* $v_u(t) + v'_u(t) \leq t/2$.

Corollary 7. *Let* $w(t) = \exp\left(\left(\frac{2\log t}{\Delta}\right)^{2/3}\right)$. *Then,*

$$\Psi_u(t) = O\left(K\frac{\log^3 t}{\epsilon_u^2 t}\right)$$

for all t such that $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$, $\frac{t}{w(t)} \geq \max\left\{\frac{24K}{\epsilon_u}, 15K^2 \log t\right\}$, *and* $\frac{t}{\log t} \geq \frac{198}{\epsilon_u^2}$.

As shown in Algorithm 2, $\mathsf{E}(t)$ indicates whether Q-ThS(match) chooses to explore at time t. We now obtain a bound on the expected number of time-slots Q-ThS(match) chooses to explore in an arbitrary time interval $(t_1, t_2]$. Since at any time t, Q-ThS(match) decides to explore with probability $\min\{1, 3K\frac{\log^2 t}{t}\}$, we have

$$\mathbb{E}\left[\sum_{l=t_1+1}^{t_2} \mathsf{E}(l)\right] \leq 3K \sum_{l=t_1+1}^{t_2} \frac{\log^2 l}{l} \leq 3K \int_{t_1}^{t_2} \frac{\log^2 l}{l}\, \mathrm{d}l = K\left(\log^3 t_2 - \log^3 t_1\right). \tag{11}$$

The following lemma gives a probabilistic upper bound on the same quantity.

Lemma 8. *For any t and $t_1 < t_2$,*

$$\mathbb{P}\left[\sum_{l=t_1+1}^{t_2} \mathsf{E}(l) \geq 5\max\left(\log t, K\left(\log^3 t_2 - \log^3 t_1\right)\right)\right] \leq \frac{1}{t^4}.$$

Proof. To prove the result, we will use the following Chernoff bound: for a sum of independent Bernoulli random variables Y with mean $\mathbb{E}Y$ and for any $\delta > 0$,

$$\mathbb{P}\left[[Y \geq (1+\delta)\mathbb{E}Y\right] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}Y}.$$

If $\mathbb{E}Y \geq \log t$, the above bound for $\delta = 4$ gives

$$\mathbb{P}\left[Y \geq 5\mathbb{E}Y\right] \leq \frac{1}{t^4}.$$

Note that $\{\mathsf{E}(l)\}_{l=t_1+1}^{t_2}$ are independent Bernoulli random variables and let $X = \sum_{l=t_1}^{t_2} \mathsf{E}(l)$. Now consider the probability $\mathbb{P}\left[X \geq 5\max\left(\log t, \mathbb{E}X\right)\right]$. If $\mathbb{E}X \geq \log t$, then the result is true from the above Chernoff bound. If $\mathbb{E}X < \log t$, then it is possible to construct a random variable Y which is a sum of independent Bernoulli random variables, has mean $\log t$ and stochastically dominates X, in which case we can again use the Chernoff bound on Y. Therefore,

$$\mathbb{P}\left[X \geq 5\log t\right] \leq \mathbb{P}\left[Y \geq 5\log t\right] \leq \frac{1}{t^4}.$$

Using inequality (11), we have the required result, i.e.,

$$\mathbb{P}\left[\sum_{l=t_1+1}^{t_2} \mathsf{E}(l) \geq 5\max\left(\log t, K\left(\log^3 t_2 - \log^3 t_1\right)\right)\right] \leq \mathbb{P}\left[X \geq 5\max\left(\log t, \mathbb{E}X\right)\right] \leq 1/t^4.$$

□

Let $w(t) = \exp\left(\left(\frac{2\log t}{\Delta}\right)^{2/3}\right)$. The next lemma shows that, with high probability, Q-ThS(match) does not schedule a sub-optimal matching when it exploits in the late stage.

Lemma 9. *For $t \geq \exp\left(6/\Delta^2\right)$,*

$$\mathbb{P}\left[\bigcup_{u\in[U]} \sum_{l=w(t)+1}^{t} \sum_{k\neq k_u^*} \mathsf{I}_{uk}(l) > 0\right] = O\left(\frac{UK}{t^3}\right).$$

Proof. Let $X_{uk}(l), u = 1, 2, .., U, k = 1, 2, .., K, l = 1, 2, 3..$ be independent random variables denoting the service offered in the l^{th} assignment of the server k to queue u. Consider the events,

$$T_{uk}(w(t)) \geq \frac{1}{2}\log^3(w(t)),\ \forall k \in [K], u \in [U] \tag{12}$$

$$\theta_{uk_u^*}(s) > \mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}},\ \forall s,\ \text{s.t. } w(t)+1 \le s \le t, u \in [U] \tag{13}$$

and

$$\theta_{uk}(s) \le \mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}},\ \forall s, k \text{ s.t. } w(t)+1 \le s \le t, k \neq k_u^*, u \in [U] \tag{14}$$

It can be seen that, given the above events, Q-ThS(match) schedules the optimal matching in all time-slots in $(w(t), t]$ in which it decides to exploit, i.e., $\sum_{l=w(t)+1}^{t} \sum_{k \neq k_u^*} \mathsf{I}_{uk}(l) = 0$ for all $u \in [U]$. We now show that the events above occur with high probability.

Note that, since the matchings in $\mathcal{E}$ cover all the links in the system, $T_{uk}(w(t)) \le \frac{1}{2}\log^3(w(t))$ for some u, k implies that $\sum_{l=1}^{w(t)} \mathbb{1}\{\boldsymbol{\kappa}(t) = \boldsymbol{\kappa}\} \le \frac{1}{2}\log^3(w(t))$ for some $\boldsymbol{\kappa} \in \mathcal{E}$. Since $\sum_{l=1}^{w(t)} \mathbb{1}\{\boldsymbol{\kappa}(t) = \boldsymbol{\kappa}\}$ is a sum of i.i.d. Bernoulli random variables with mean $\log^3(w(t))$, we use Chernoff bound to prove that event (12) occurs with high probability.

$$\begin{aligned}
\mathbb{P}\left[(12) \text{ is false}\right] &\le \sum_{\boldsymbol{\kappa} \in \mathcal{E}} \mathbb{P}\left[\sum_{l=1}^{w(t)} \mathbb{1}\{\boldsymbol{\kappa}(t) = \boldsymbol{\kappa}\} \le \frac{1}{2}\log^3(w(t))\right] \\
&\le K \exp\left(-\frac{1}{8}\log^3(w(t))\right) \\
&= K \exp\left(-\frac{1}{8}\left(\frac{2\log t}{\Delta}\right)^2\right) = o\left(\frac{K}{t^4}\right).
\end{aligned} \tag{15}$$

In order to prove high probability bounds for the other two events, we define U_s to be a sequence of i.i.d uniform random variables taking values in $[0,1]$ for $s = w(t)+1, ..., t$. Let us also define $\Sigma_{u,k,l} = \sum_{r=1}^{l} X_{uk}(r)$. In what follows let $F_{a,b}^{\text{Beta}}$ denote the c.d.f of the $\text{Beta}(a,b)$ distribution while $F_{n,p}^{\text{B}}$ denotes the c.d.f. of a $\text{Binomial}(n,p)$ distribution. Let $S_{uk}(t) = m\hat{i}u_{uk}(t)T_{uk}(t)$ for all $u \in [U], k \in [K]$.

$$\begin{aligned}
\mathbb{P}\left[(13) \text{ is false}\right] &\le \sum_{u \in [U]} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\theta_{uk_u^*}(s) \le \mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}}\right] \\
&= \sum_{u \in [U]} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[U_s \le F^{\text{Beta}}_{S_{uk_u^*}(s)+1, T_{uk_u^*}(s) - S_{uk_u^*}(s)+1}\left(\mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}}\right)\right] \\
&\overset{(i)}{\le} \sum_{u \in [U]} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\exists l \in \left\{\frac{1}{2}\log^3(s), ..., s\right\} : F^{B}_{l+1, \mu_u^* - \sqrt{\frac{\log^2(s)}{l}}}\left(\Sigma_{u,k_u^*,l}\right) \le U_s \middle| (12) \text{ is true}\right] \\
&+ o\left(\frac{UK}{t^3}\right) \\
&\le \sum_{u \in [U]} \sum_{s=w(t)+1}^{t} \sum_{l=\frac{1}{2}\log^3(s)}^{s} \mathbb{P}\left[\Sigma_{u,k_u^*,l} \le (F^B)^{-1}_{l+1, \mu_u^* - \sqrt{\frac{\log^2(s)}{l}}}(U_s)\right] + o\left(\frac{UK}{t^3}\right)
\end{aligned}$$

In (i) we use the well-known Beta-Binomial trick [] and the fact that given (12) is true, uk_u^* has been scheduled enough number of times. Now the term $(F^B)^{-1}_{l+1, \mu_u^* - \sqrt{\frac{\log^2(s)}{l}}}(U_s)$ can be thought of as the sum of $l+1$ i.i.d Bernoulli random variables with mean $\mu_u^* - \sqrt{\frac{\log^2(s)}{l}}$. Let Z_r be a sequence of

i.i.d random variable with mean $\sqrt{\frac{\log^2(s)}{l}}$. Therefore we have,

$$\mathbb{P}\left[\Sigma_{u,k,l} \leq (F^B)^{-1}_{l+1,\mu_u^*-\sqrt{\frac{\log^2(s)}{l}}}(U_s)\right] \leq \mathbb{P}\left[\sum_{r=1}^{l} Z_r \leq 1\right] \overset{(ii)}{\leq} e^{-\frac{\log^2(s)}{3}} \tag{16}$$

Here, (ii) is due to Chernoff-Hoeffding's inequality. Therefore we have,

$$\begin{aligned}
\mathbb{P}\left[(13) \text{ is false}\right] &\leq U \sum_{s=w(t)+1}^{t} \sum_{l=\frac{1}{2}\log^3(s)}^{s} \exp\left(-\frac{\log^2(s)}{3}\right) + o\left(\frac{UK}{t^3}\right) \\
&\leq U \exp\left(-\frac{1}{3}\log^2(w(t)) + 2\log t\right) + o\left(\frac{UK}{t^3}\right) \\
&= U \exp\left(-\frac{1}{3}\left(\frac{2\log t}{\Delta}\right)^{4/3} + 2\log t\right) + o\left(\frac{UK}{t^3}\right) = o\left(\frac{UK}{t^3}\right).
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\left[(14) \text{ is false}\right] &\leq \sum_{u\in[U],k\neq k_u^*} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\theta_{uk}(s) > \mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}}\right] \\
&\leq \sum_{u\in[U],k\neq k_u^*} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\theta_{uk}(s) > \mu_u^* - \sqrt{\frac{\log^2(s)}{T_{uk_u^*}(s)}} \middle| (12) \text{ is true}\right] + o\left(\frac{UK}{t^3}\right) \\
&\overset{(iii)}{\leq} \sum_{u\in[U],k\neq k_u^*} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\theta_{uk}(s) > \mu_u^* - \sqrt{\frac{2}{\log(s)}} \middle| (12) \text{ is true}\right] + o\left(\frac{UK}{t^3}\right) \\
&\overset{(iv)}{\leq} \sum_{u\in[U],k\neq k_u^*} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\theta_{uk}(s) > \mu_{uk} + \frac{\Delta}{2} \middle| (12) \text{ is true}\right] + o\left(\frac{UK}{t^3}\right) \\
&\overset{(v)}{\leq} \sum_{u\in[U],k\neq k_u^*} \sum_{s=w(t)+1}^{t} \mathbb{P}\left[\exists l \in \left\{\frac{1}{2}\log^3(s), ..., s\right\} : \Sigma_{u,k,l} \geq (F^B)^{-1}_{l+1,\mu_{uk}+\frac{\Delta}{2}}(U_s)\right] + o\left(\frac{UK}{t^3}\right) \\
&\overset{(vi)}{\leq} o\left(\frac{UK}{t^3}\right)
\end{aligned}$$

We observe that given (12) is true, we have scheduled uk_u^* enough number of times in order to get (iii). In (iv) we use that fact that $t \geq \exp\left(6/\Delta^2\right)$. (v) is due to the Beta-Binomial trick while (vi) is a result of applying the Chernoff-Hoeffding bound to the first term in (v) in a manner similar to that of (16). □

For any time t, let

$$B_u(t) := \min\{s \geq 0 : Q_u(t-s) = 0\}$$

denote the time elapsed since the beginning of the current regenerative cycle for queue u. Alternately, at any time t, $t - B_u(t)$ is the last time instant at which queue u was zero.

The following lemma gives an upper bound on the sample-path queue-regret in terms of the number of sub-optimal schedules in the current regenerative cycle.

Lemma 10. *For any $t \geq 1$,*

$$Q_u(t) - Q_u^*(t) \leq \sum_{l=t-B_u(t)+1}^{t} \left(\mathsf{E}(l) + \sum_{k\neq k_u^*} \mathsf{I}_{uk}(l)\right).$$

Proof. If $B_u(t) = 0$, i.e., if $Q_u(t) = 0$, then the result is trivially true.

Consider the case where $B_u(t) > 0$. Since $Q_u(l) > 0$ for all $t - B_u(t) + 1 \leq l \leq t$, we have

$$Q_u(l) = Q_u(l-1) + A_u(l) - S_u(l) \quad \forall t - B_u(t) + 1 \leq l \leq t.$$

This implies that

$$Q_u(t) = \sum_{l=t-B_u(t)+1}^{t} A_u(l) - S_u(l).$$

Moreover,

$$Q_u^*(t) = \max_{1 \leq s \leq t} \left(Q_u^*(0) + \sum_{l=s}^{t} A_u(l) - S_u^*(l) \right)^+ \geq \sum_{l=t-B_u(t)+1}^{t} A_u(l) - S_u^*(l).$$

Combining the above two expressions, we have

$$\begin{aligned}
Q_u(t) - Q_u^*(t) &\leq \sum_{l=t-B_u(t)+1}^{t} S_u^*(l) - S_u(l) \\
&= \sum_{l=t-B_u(t)+1}^{t} \sum_{k \in [K]} \left(R_{uk_u^*}(l) - R_{uk}(l) \right) \left(\mathsf{E}_{uk}(l) + \mathsf{I}_{uk}(l) \right) \\
&\leq \sum_{l=t-B_u(t)+1}^{t} \sum_{k \neq k_u^*} \left(\mathsf{E}_{uk}(l) + \mathsf{I}_{uk}(l) \right) \\
&\leq \sum_{l=t-B_u(t)+1}^{t} \left(\mathsf{E}(l) + \sum_{k \neq k_u^*} \mathsf{I}_{uk}(l) \right),
\end{aligned}$$

where the second inequality follows from the assumption that the service provided by each of the links is bounded by 1, and the last inequality from the fact that $\sum_{k \in [K]} \mathsf{E}_{uk}(l) = \mathsf{E}(l)\ \forall l, \forall u \in [U]$. □

In the next lemma, we derive a coarse high probability upper bound on the queue-length. This bound on the queue-length is used later to obtain a first cut bound on the length of the regenerative cycle in Lemma 12.

Lemma 11. *For any* $l \in [1, t]$,

$$\mathbb{P}\left[Q_u(l) > 2Kw(t) \right] = O\left(\frac{UK}{t^3} \right)$$

$\forall t$ *s.t.* $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$ *and* $t \geq \exp\left(6/\Delta^2\right)$.

Proof. From Lemma 10,

$$Q_u(t) - Q_u^*(t) \leq \sum_{l=t-B_u(t)+1}^{t} \left(\mathsf{E}(l) + \sum_{k \neq k_u^*} \mathsf{I}_{uk}(l) \right) \leq \sum_{l=1}^{t} \left(\mathsf{E}(l) + \sum_{k \neq k_u^*} \mathsf{I}_{uk}(l) \right).$$

Since $Q_u^*(t)$ is distributed according to $\pi_{(\lambda_u, \mu_u^*)}$,

$$\mathbb{P}\left[Q_u^*(t) > w(t) \right] = \frac{\lambda_u}{\mu_u^*} \left(\frac{\lambda_u \left(1 - \mu_u^*\right)}{\left(1 - \lambda_u\right) \mu_u^*} \right)^{w(t)} \leq \exp\left(w(t) \log \left(\frac{\lambda_u \left(1 - \mu_u^*\right)}{\left(1 - \lambda_u\right) \mu_u^*} \right) \right) \leq \frac{1}{t^3}$$

if $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$. The last inequality follows from the following bound –

$$\begin{aligned}
\log \left(\frac{\left(1 - \lambda_u\right) \mu_u^*}{\lambda_u \left(1 - \mu_u^*\right)} \right) &= \log \left(1 + \frac{\epsilon_u}{\lambda_u \left(1 - \mu_u^*\right)} \right) \\
&\geq \log\left(1 + 4\epsilon_u\right) \quad \text{since } \left(\lambda_u \left(1 - \mu_u^*\right) < 1/4\right) \\
&\geq \frac{3}{2} \epsilon_u.
\end{aligned}$$

Moreover, from Lemma 8, we have

$$\mathbb{P}\left[\sum_{l=1}^{t}\mathsf{E}(l) > Kw(t)\right] = o\left(\frac{1}{t^3}\right).$$

Now, note that

$$\sum_{l=1}^{t}\sum_{k\neq k_u^*}\mathsf{I}_{uk}(l) \leq (K-1)w(t) + \sum_{l=w(t)+1}^{t}\sum_{k\neq k_u^*}\mathsf{I}_{uk}(l).$$

Therefore,

$$\mathbb{P}\left[\sum_{l=1}^{t}\sum_{k\neq k_u^*}\mathsf{I}_{uk}(l) > (K-1)w(t)\right] \leq \mathbb{P}\left[\sum_{l=w(t)+1}^{t}\sum_{k\neq k_u^*}\mathsf{I}_{uk}(l) > 0\right] = O\left(\frac{UK}{t^3}\right)$$

from Lemma 9. Using the inequalities above, we have

$$\begin{aligned}\mathbb{P}\left[Q_u(t) > 2Kw(t)\right] &\leq \mathbb{P}\left[Q_u^*(t) > w(t)\right] + \mathbb{P}\left[\sum_{l=1}^{t}\mathsf{E}(l) > Kw(t)\right] \\ &\quad + \mathbb{P}\left[\sum_{l=1}^{t}\sum_{k\neq k_u^*}\mathsf{I}_{uk}(l) > (K-1)w(t)\right] \\ &\leq \frac{1}{t^3} + O\left(\frac{UK}{t^3}\right) \\ &= O\left(\frac{UK}{t^3}\right).\end{aligned}$$

□

Lemma 12. *Let $v_u'(t) = \frac{6K}{\epsilon_u}w(t)$ and let v_u be an arbitrary function. Then,*

$$\mathbb{P}\left[B_u\left(t - v_u(t)\right) > v_u'(t)\right] = O\left(\frac{UK}{t^3}\right)$$

$\forall t$ s.t. $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}, t \geq \exp\left(6/\Delta^2\right)$ and $v_u(t) + v_u'(t) \leq t/2$.

Proof. Let $r(t) := t - v_u(t)$. Consider the events

$$Q_u(r(t) - v_u'(t)) \leq 2Kw(t), \tag{17}$$

$$\sum_{l=r(t)-v_u'(t)+1}^{r(t)} A_u(l) - R_{uk_u^*}(l) \leq -\frac{\epsilon_u}{2}v_u'(t), \tag{18}$$

$$\sum_{l=r(t)-v_u'(t)+1}^{r(t)}\mathsf{E}(l) + \sum_{k\neq k_u^*}\mathsf{I}_{uk}(l) \leq Kw(t). \tag{19}$$

By the definition of $v_u'(t)$,

$$2Kw(t) - \frac{\epsilon_u}{2}v_u'(t) \leq -Kw(t).$$

Given Events (17)-(19), the above inequality implies that

$$\begin{aligned}Q_u(r(t) - v_u'(t)) + \sum_{l=r(t)-v_u'(t)+1}^{r(t)} A_u(l) &\leq \sum_{l=r(t)-v_u'(t)+1}^{r(t)} R_{uk_u^*}(l) - \left(\mathsf{E}(l) + \sum_{k\neq k_u^*}\mathsf{I}_{uk}(l)\right) \\ &\leq \sum_{l=r(t)-v_u'(t)+1}^{r(t)} S_u(l),\end{aligned}$$

which further implies that $Q_u(l) = 0$ for some $l \in [r(t) - v'_u(t) + 1, r(t)]$. This gives us that $B_u(r(t)) \leq v'_u(t)$.

We now show that each of the events (17)-(19) occur with high probability. Consider the event (18) and note that $A_u(l) - R_{uk^*_u}(l)$ are i.i.d. random variables with mean $-\epsilon_u$ and bounded between -1 and 1. Using Chernoff bound for sum of bounded i.i.d. random variables, we have

$$\mathbb{P}\left[\sum_{l=r(t)-v'_u(t)+1}^{r(t)} A_u(l) - R_{uk^*_u}(l) > -\frac{\epsilon_u}{2} v'_u(t)\right] \leq \exp\left(-\frac{\epsilon_u^2}{8} v'_u(t)\right) \leq \frac{1}{t^3}$$

since $v'_u(t) \geq \frac{6K}{\epsilon_u} w(t) \geq \frac{24}{\epsilon_u^2} \log t$.

By Lemmas 11, 9 and 8, the probability that any of the events (17), (19) does not occur is $O\left(\frac{UK}{t^3}\right)$ $\forall t$ s.t. $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$ and $v_u(t) + v'_u(t) \leq t/2$, and therefore we have the required result. □

Using the preceding upper bound on the regenerative cycle-length, we derive tighter bounds on the queue-length and the regenerative cycle-length in Lemmas 14 and 15 respectively. The following lemma is a useful intermediate result.

Lemma 13. *For any $u \in [U]$ and t_2 s.t. $1 \leq t_2 \leq t$,*

$$\mathbb{P}\left[\max_{1 \leq s \leq t_2} \left\{\sum_{l=t_2-s+1}^{t_2} A_u(l) - R_{uk^*_u}(l)\right\} \geq \frac{2 \log t}{\epsilon_u}\right] \leq \frac{1}{t^3}.$$

Proof. Let $X_s = \sum_{l=t_2-s+1}^{t_2} A_u(l) - R_{uk^*_u}(l)$. Since X_s is the sum of s i.i.d. random variables with mean ϵ_u and is bounded within $[-1, 1]$, Hoeffding's inequality gives

$$\begin{aligned}
\mathbb{P}\left[X_s \geq \frac{2 \log t}{\epsilon_u}\right] &= \mathbb{P}\left[X_s - \mathbb{E}X_s \geq \epsilon_u s + \frac{2 \log t}{\epsilon_u}\right] \\
&\leq \exp\left(-\frac{2\left(\epsilon_u s + \frac{2 \log t}{\epsilon_u}\right)^2}{4s}\right) \\
&\leq \exp\left(-4 \log t\right),
\end{aligned}$$

where the last inequality follows from the fact that $(a+b)^2 > 4ab$ for any $a, b \geq 0$. Using union bound over all $1 \leq s \leq t_2$ gives the required result. □

Lemma 14. *Let $v'_u(t) = \frac{6K}{\epsilon_u} w(t)$ and v_u be an arbitrary function. Then,*

$$\mathbb{P}\left[Q_u(t - v_u(t)) > \left(\frac{2}{\epsilon_u} + 5\right) \log t + 30K \frac{v'_u(t) \log^2 t}{t}\right] = O\left(\frac{UK}{t^3}\right)$$

$\forall t$ s.t. $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$, $t \geq \exp\left(6/\Delta^2\right)$ and $v_u(t) + v'_u(t) \leq t/2$.

Proof. Let $r(t) = t - v_u(t)$. Now, consider the events

$$B_u(r(t)) \leq v'_u(t), \tag{20}$$

$$\sum_{l=r(t)-s+1}^{r(t)} A_u(l) - R_{uk^*_u}(l) \leq \frac{2 \log t}{\epsilon_u} \; 1 \leq s \leq v'_u(t), \tag{21}$$

$$\sum_{l=r(t)-v'_u(t)+1}^{r(t)} \mathsf{E}(l) + \sum_{k \neq k^*_u} \mathsf{I}_{uk}(l) \leq 5 \log t + 5K\left(\log^3\left(r(t)\right) - \log^3\left(r(t) - v'_u(t)\right)\right). \tag{22}$$

Given the above events, we have

$$\begin{aligned}
Q_u(r(t)) &= \sum_{l=r(t)-B_u(r(t))+1}^{r(t)} A_u(l) - S(l) \\
&\leq \sum_{l=r(t)-B_u(r(t))+1}^{r(t)} A_u(l) - R_{uk_u^*}(l) + \mathsf{E}(l) + \sum_{k\neq k^*} \mathsf{I}_{uk}(l) \\
&\leq \left(\frac{2}{\epsilon_u}+5\right)\log t + 5K\left(\log^3\left(r(t)\right) - \log^3\left(r(t)-v_u'(t)\right)\right) \\
&\leq \left(\frac{2}{\epsilon_u}+5\right)\log t + 15K\frac{v_u'(t)\log^2 t}{(r(t)-v_u'(t))} \\
&\leq \left(\frac{2}{\epsilon_u}+5\right)\log t + 30K\frac{v_u'(t)\log^2 t}{t},
\end{aligned}$$

where the last inequality is true if $v_u(t) + v_u'(t) \leq t/2$. From Lemmas 12, 13, 9 and 8, probability of each the events (20)-(22) is $1 - O\left(\frac{UK}{t^3}\right)$ and therefore, we have the required result. □

Lemma 15. *Let* $v_u'(t) = \frac{6K}{\epsilon_u} w(t)$ *and* $v_u(t) = \frac{24\log t}{\epsilon_u^2} + \frac{60K}{\epsilon_u}\frac{v_u'(t)\log^2 t}{t}$. *Then,*

$$\mathbb{P}\left[B_u(t) > v_u(t)\right] = O\left(\frac{UK}{t^3}\right)$$

$\forall t$ *s.t.* $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$,$t \geq \exp\left(6/\Delta^2\right)$ *and* $v_u(t) + v_u'(t) \leq t/2$.

Proof. Let $r(t) = t - v_u(t)$. As in Lemma 12, consider the events

$$Q_u(r(t)) \leq \left(\frac{2}{\epsilon_u}+5\right)\log t + 30K\frac{v_u'(t)\log^2 t}{t}, \tag{23}$$

$$\sum_{l=r(t)+1}^{t} A_u(l) - R_{uk_u^*}(l) \leq -\frac{\epsilon_u}{2} v_u(t), \tag{24}$$

$$\sum_{l=r(t)+1}^{t} \mathsf{E}(l) + \sum_{k\neq k_u^*} \mathsf{I}_{uk}(l) \leq 5\log t + 5K\left(\log^3 t - \log^3\left(r(t)\right)\right). \tag{25}$$

The definition of $v_u(t)$ and events (23)-(25) imply that

$$\begin{aligned}
Q_u(r(t)) + \sum_{l=r(t)+1}^{t} A_u(l) &\leq \sum_{l=r(t)+1}^{t} R_{uk_u^*}(l) - \sum_{l=r(t)+1}^{t} \mathsf{E}(l) + \sum_{k\neq k_u^*} \mathsf{I}_{uk}(l) \\
&\leq \sum_{l=r(t)+1}^{t} S_u(l),
\end{aligned}$$

which further implies that $Q(l) = 0$ for some $l \in [r(t)+1, t]$ and therefore $B_u(t) \leq v_u(t)$. We can again show that each of the events (23)-(25) occurs with high probability. Particularly, by Lemmas 8, 9 and 14, the probability that any one of the events (23), (25) does not occur is $O\left(\frac{UK}{t^3}\right)$ $\forall t$ s.t. $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$ and $v_u(t) + v_u'(t) \leq t/2$. We can bound the probability of event (24) in the same way as event (21) in Lemma 12 to show that it occurs with probability at least $\frac{1}{t^3}$. Combining all these gives us the required high probability result. □

Proof of Theorem 6. The proof is based on two main ideas: one is that the regenerative cycle length is not very large, and the other is that the algorithm has correctly identified the optimal matching

in late stages. We combine Lemmas 9 and 15 to bound the regret at any time t s.t. $\frac{w(t)}{\log t} \geq \frac{2}{\epsilon_u}$ and $v_u(t) + v'_u(t) \leq t/2$:

$$
\begin{aligned}
\Psi_u(t) &= \mathbb{E}\left[Q_u(t) - Q^*_u(t)\right] \\
&\leq \mathbb{E}\left[Q_u(t) - Q^*_u(t) \middle| B_u(t) \leq v_u(t)\right] \mathbb{P}\left[B_u(t) \leq v_u(t)\right] \\
&\quad + \mathbb{E}\left[Q_u(t) - Q^*_u(t) \middle| B_u(t) > v_u(t)\right] \mathbb{P}\left[B_u(t) > v_u(t)\right] \\
&\leq \mathbb{E}\left[\sum_{l=t-v_u(t)+1}^{t} \mathsf{E}(l) + \sum_{k \neq k^*_u} \mathsf{I}_{uk}(l)\right] + t\mathbb{P}\left[B_u(t) > v_u(t)\right] \qquad (26) \\
&\leq K\left(\log^3(t) - \log^3(t - v_u(t))\right) + t\mathbb{P}\left[\sum_{l=t-v_u(t)+1}^{t} \sum_{k \neq k^*_u} \mathsf{I}_{uk}(l) > 0\right] + t\mathbb{P}\left[B_u(t) > v_u(t)\right] \qquad (27) \\
&\leq 3K \log^2 t \log\left(1 + \frac{v_u(t)}{t - v_u(t)}\right) + O\left(\frac{UK}{t^2}\right) \\
&= O\left(K \frac{v_u(t) \log^2 t}{t - v_u(t)}\right) + O\left(\frac{U}{tw(t)}\right) \\
&= O\left(K \frac{v_u(t) \log^2 t}{t}\right),
\end{aligned}
$$

where (26) follows from Lemma 10, and the last two terms in inequality (27) are bounded using Lemmas 9 and 15. □

Proof of Corollary 7. We first note the following:

(i) $\frac{t}{w(t)} \geq \frac{24K}{\epsilon_u}$ implies that $v'_u(t) \leq \frac{t}{4}$,

(ii) $\frac{t}{w(t)} \geq 15K^2 \log t$ implies that $\frac{24}{\epsilon_u^2} \log t \geq \frac{60K}{\epsilon_u} \frac{v'_u(t) \log^2 t}{t}$, and therefore $v_u(t) \leq \frac{48}{\epsilon_u^2} \log t$

(iii) $\frac{t}{\log t} \geq \frac{198}{\epsilon_u^2}$ implies that $v_u(t) \leq \frac{t}{4}$.

These inequalities when applied to Theorem 6 give the required result. □

8.2 Lower Bounds for α-Consistent Policies

As mentioned earlier, we prove asymptotic and early stage lower bounds for a class of policies called the α-consistent class (Definition 1). As before we will be proving our results for a more general case where there are U queues and K servers. Theorems 1 and 4 are special cases of the analogous theorems stated below, under the unique optimal matching assumption.

Theorem 16. *For any problem instance $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ with a unique optimal matching, and any α-consistent policy, the regret $\boldsymbol{\Psi}(t)$ satisfies*

(a)

$$\frac{1}{U} \sum_{u \in [U]} \Psi_u(t) \geq \left(\frac{\lambda_{min}}{8} D(\boldsymbol{\mu})(1-\alpha)(K-1)\right) \frac{1}{t},$$

(b) and for any $u \in [U]$,

$$\Psi_u(t) \geq \left(\frac{\lambda_{min}}{8} D(\boldsymbol{\mu})(1-\alpha) \max\{U-1, 2(K-U)\}\right) \frac{1}{t}$$

for infinitely many t, where

$$D(\boldsymbol{\mu}) = \frac{\Delta}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)}. \tag{28}$$

Theorem 17. *Given any problem instance $(\boldsymbol{\lambda}, \boldsymbol{\mu})$, and for any α-consistent policy and $\gamma > \frac{1}{1-\alpha}$, the regret $\boldsymbol{\Psi}(t)$ satisfies*

(a)

$$\frac{1}{U}\sum_{u\in[U]} \Psi_u(t) \geq \frac{D(\boldsymbol{\mu})}{4}(K-1)\frac{\log t}{\log\log t},$$

for $t \in \left[\max\{C_4 K^\gamma, \tau\}, (K-1)\frac{D(\boldsymbol{\mu})}{4\bar{\epsilon}}\right]$, *and*

(b) for any $u \in [U]$,

$$\Psi_u(t) \geq \frac{D(\boldsymbol{\mu})}{4}\max\{U-1, 2(K-U)\}\frac{\log t}{\log\log t}$$

for $t \in \left[\max\{C_4 K^\gamma, \tau\}, (K-1)\frac{D(\boldsymbol{\mu})}{2\epsilon_u}\right]$,

where $D(\boldsymbol{\mu})$ is given by equation 28, $\bar{\epsilon} = \frac{1}{U}\sum_{u\in[U]}\epsilon_u$, and τ and C_4 are constants that depend on α, γ and the policy.

In order to prove Theorems 16 and 17, we use techniques from existing work in the MAB literature along with some new lower bounding ideas specific to queueing systems. Specifically, we use lower bounds for α-consistent policies on the expected number of times a sub-optimal server is scheduled. This lower bound, shown (in Lemma 19) specifically for the problem of scheduling a unique optimal matching, is similar in style to the traditional bandit lower bound by Lai et al. [7] but holds in the non-asymptotic setting. Also, as opposed the traditional change of measure proof technique used in [7], the proof (similar to the more recent ones [20, 21, 18]) uses results from hypothesis testing (Lemma 18).

Lemma 18 ([22])**.** *Consider two probability measures P and Q, both absolutely continuous with respect to a given measure. Then for any event $\mathcal{A}$ we have:*

$$P(\mathcal{A}) + Q(\mathcal{A}^c) \geq \frac{1}{2}\exp\{-\min(\mathrm{KL}(P||Q), \mathrm{KL}(Q||P))\}.$$

Proof. Let $p = P(\mathcal{A})$ and $q = Q(\mathcal{A}^c)$. From standard properties of KL divergence we have that,

$$\mathrm{KL}(P||Q) \geq \mathrm{KL}(p, q)$$

Therefore, it is sufficient to prove that

$$p + q \geq \frac{1}{2}\exp\left(-p\log\frac{p}{1-q} - (1-p)\log\frac{1-p}{q}\right) = \frac{1}{2}\left(\frac{1-q}{p}\right)^p\left(\frac{q}{1-p}\right)^{1-p}.$$

Now,

$$\begin{aligned}
\left(\frac{1-q}{p}\right)^p\left(\frac{q}{1-p}\right)^{1-p} &= \left(\sqrt{\frac{1-q}{p}}\right)^{2p}\left(\sqrt{\frac{q}{1-p}}\right)^{2(1-p)} \\
&\leq \left(\frac{1}{2}\left(2p\cdot\sqrt{\frac{1-q}{p}} + 2(1-p)\cdot\sqrt{\frac{q}{1-p}}\right)\right)^2 \\
&= \left(\sqrt{p(1-q)} + \sqrt{q(1-p)}\right)^2 \\
&\leq 2(p(1-q) + q(1-p)) \\
&< 2(p+q)
\end{aligned}$$

as required. □

Lemma 19. *For any problem instance $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and any α-consistent policy, there exist constants τ and C s.t. for any $u \in [U]$, $k \neq k_u^*$ and $t > \tau$,*

$$\mathbb{E}\left[T_{uk}(t)\right] + \sum_{u' \neq u} \mathbb{1}\left\{k_{u'}^* = k\right\} \mathbb{E}\left[T_{u'k_u^*}(t)\right] \geq \frac{1}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)} \left((1-\alpha)\log t - \log(4KC)\right).$$

Proof. Without loss of generality, let the optimal servers for the U queues be denoted by the first U indices. In other words, a server $k > U$ is not an optimal server for any queue, i.e., for any $u' \in [U]$, $K \geq k > U$, $\mathbb{1}\left\{k_{u'}^* = k\right\} = 0$. Also, let $\beta = \frac{\mu_{max}+1}{2}$.

We will first consider the case $k \leq U$. For a fixed user u and server $k \leq U$, let u' be the queue that has k as the best server, i.e., $k_{u'}^* = k$. Now consider the two problem instances $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and $(\boldsymbol{\lambda}, \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ is the same as $\boldsymbol{\mu}$ except for the two entries corresponding to indices $(u, k), (u', k_u^*)$ replaced by β. Therefore, for the problem instance $(\boldsymbol{\lambda}, \hat{\boldsymbol{\mu}})$, the best servers are swapped for queues u and u' and remain the same for all the other queues. Let $\mathbb{P}_{\boldsymbol{\mu}}^t$ and $\mathbb{P}_{\hat{\boldsymbol{\mu}}}^t$ be the distributions corresponding to the arrivals, chosen servers and rates obtained in the first t plays for the two instances under a fixed α-consistent policy. Recall that $T_{uk}(t) = \sum_{s=1}^t \mathbb{1}\{\kappa_u(s) = k\}\ \forall u \in [U], k \in [K]$. Define the event $\mathcal{A} = \{T_{uk}(t) > t/2\}$. By the definition of α-consistency there exists a fixed integer τ and a fixed constant C such that for all $t > \tau$ we have,

$$\mathbb{E}_{\boldsymbol{\mu}}^t \left[\sum_{s=1}^t \mathbb{1}\{\kappa_u(s) = k\}\right] \leq Ct^\alpha$$
$$\mathbb{E}_{\hat{\boldsymbol{\mu}}}^t \left[\sum_{s=1}^t \mathbb{1}\{\kappa_u(s) = k'\}\right] \leq Ct^\alpha\ , \forall k' \neq k.$$

A simple application of Markov's inequality yields

$$\mathbb{P}_{\boldsymbol{\mu}}^t(\mathcal{A}) \leq \frac{2C}{t^{1-\alpha}}$$
$$\mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\mathcal{A}^c) \leq \frac{2C(K-1)}{t^{1-\alpha}}.$$

We can now use Lemma 18 to conclude that

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t) \geq (1-\alpha)\log t - \log(4KC). \tag{29}$$

It is, therefore, sufficient to show that

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^t || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t\right) = \mathrm{KL}\left(\mu_{uk}, \beta\right) \mathbb{E}_{\boldsymbol{\mu}}^t[T_{uk}(t)] + \mathrm{KL}\left(\mu_{u'k_u^*}, \beta\right) \mathbb{E}_{\boldsymbol{\mu}}^t[T_{u'k_u^*}(t)].$$

For the sake of brevity we write the scheduling sequence in the first t time-slots $\{\boldsymbol{\kappa}(1), \boldsymbol{\kappa}(2), ..., \boldsymbol{\kappa}(t)\}$ as $\boldsymbol{\kappa}^{(t)}$, and similarly we define $\mathbf{A}^{(t)}$ as the number of arrivals to the queue and $\mathbf{S}^{(t)}$ as the service offered by the scheduled servers in the first t time-slots. Let $\mathbf{Z}^{(t)} = (\boldsymbol{\kappa}^{(t)}, \mathbf{A}^{(t)}, \mathbf{S}^{(t)})$. The KL-divergence term can now be written as

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t) = \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t(\mathbf{Z}^{(t)}) || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\mathbf{Z}^{(t)})).$$

We can apply the chain rule of divergence to conclude that

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t(\mathbf{Z}^{(t)}) || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\mathbf{Z}^{(t)})) &= \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t(\mathbf{Z}^{(t-1)}) || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\mathbf{Z}^{(t-1)})) \\ &+ \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t(\boldsymbol{\kappa}(t) \mid \mathbf{Z}^{(t-1)}) || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\boldsymbol{\kappa}(t) \mid \mathbf{Z}^{(t-1)})) \\ &+ \mathbb{E}_{\boldsymbol{\mu}}^t\left[\mathbb{1}\{\kappa_u(t) = k\}\mathrm{KL}\left(\mu_{uk}, \beta\right) + \mathbb{1}\{\kappa_{u'}(t) = k_u^*\}\mathrm{KL}\left(\mu_{u'k_u^*}, \beta\right)\right].\end{aligned}$$

We can apply this iteratively to obtain

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t) &= \sum_{s=1}^t \mathbb{E}_{\boldsymbol{\mu}}^t\left[\mathbb{1}\{\kappa_u(s) = k\}\mathrm{KL}\left(\mu_{uk}, \beta\right)\right] \\ &+ \sum_{s=1}^t \mathbb{E}_{\boldsymbol{\mu}}^t\left[\mathbb{1}\{\kappa_{u'}(s) = k_u^*\}\mathrm{KL}\left(\mu_{u'k_u^*}, \beta\right)\right] \\ &+ \sum_{l=1}^t \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}^t(\boldsymbol{\kappa}(l) \mid \mathbf{Z}^{(l-1)}) || \mathbb{P}_{\hat{\boldsymbol{\mu}}}^t(\boldsymbol{\kappa}(l) \mid \mathbf{Z}^{(l-1)}))\end{aligned} \tag{30}$$

Note that the second summation in (30) is zero, as over a sample path the policy pulls the same servers irrespective of the parameters. Therefore, we obtain

$$\mathrm{KL}(\mathbb{P}^t_{\boldsymbol{\mu}}||\mathbb{P}^t_{\hat{\boldsymbol{\mu}}}) = \mathrm{KL}\left(\mu_{uk}, \beta\right) \mathbb{E}^t_{\boldsymbol{\mu}}[T_{uk}(t)] + \mathrm{KL}\left(\mu_{u'k^*_u}, \beta\right) \mathbb{E}^t_{\boldsymbol{\mu}}[T_{u'k^*_u}(t)],$$

which can be substituted in (29) to obtain the required result for $K \leq U$.

Now, consider the case $k > U$, where $\sum_{u\in U} \mathbb{1}\left\{k^*_u = k\right\} = 0$. We again compare the two problem instances $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and $(\boldsymbol{\lambda}, \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ is the same as $\boldsymbol{\mu}$ except for the entry corresponding to index (u, k) replaced by β. Therefore, for the problem instance $(\boldsymbol{\lambda}, \hat{\boldsymbol{\mu}})$, the best server for user u is server k while the best servers for all other queues remain the same. We can again use the same technique as before to obtain

$$\mathrm{KL}(\mathbb{P}^t_{\boldsymbol{\mu}}||\mathbb{P}^t_{\hat{\boldsymbol{\mu}}}) = \mathrm{KL}\left(\mu_{uk}, \beta\right) \mathbb{E}^t_{\boldsymbol{\mu}}[T_{uk}(t)],$$

which, along with (29), gives the required result for $K > U$. □

As a corollary of the above result, we now derive lower bound on the total expected number of sub-optimal schedules summed across all queues. In addition, we also show, for each individual queue, a lower bound for those servers which are sub-optimal for all the queues. As in the proof of Lemma 19, we assume without loss of generality that the first U indices denote the optimal servers for the U queues.

Corollary 20. *For any problem instance $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and any α-consistent policy, there exist constants τ and C s.t. for any $t > \tau$,*

(a)

$$2\Delta \sum_{u\in[U]} \sum_{k\neq k^*_u} \mathbb{E}\left[T_{uk}(t)\right] \geq U(K-1)D(\boldsymbol{\mu})\left((1-\alpha)\log t - \log(4KC)\right),$$

(b) for any $u \in [U]$,

$$2\Delta \sum_{k\neq k^*_u} \mathbb{E}\left[T_{uk}(t)\right] \geq (U-1)D(\boldsymbol{\mu})\left((1-\alpha)\log t - \log(4KC)\right),$$

(c) and for any $u \in [U]$,

$$\Delta \sum_{k>U} \mathbb{E}\left[T_{uk}(t)\right] \geq (K-U)D(\boldsymbol{\mu})\left((1-\alpha)\log t - \log(4KC)\right),$$

where $D(\boldsymbol{\mu})$ is given by (28).

Proof. To prove part (a), we observe that a unique optimal server for each queue in the system implies that

$$\begin{aligned}\sum_{u\in[U]} \sum_{k\neq k^*_u} \mathbb{E}\left[T_{uk}(t)\right] &\geq \sum_{u\in[U]} \sum_{u'\neq u} \mathbb{E}\left[T_{uk^*_{u'}}(t)\right] \\ &= \sum_{u\in[U]} \sum_{k\neq k^*_u} \sum_{u'\neq u} \mathbb{1}\left\{k^*_{u'} = k\right\} \mathbb{E}\left[T_{u'k^*_u}(t)\right].\end{aligned}$$

Now, from Lemma 19, there exist constants C and τ such that for $t > \tau$,

$$\begin{aligned}2\sum_{u\in[U]} \sum_{k\neq k^*_u} \mathbb{E}\left[T_{uk}(t)\right] &\geq \sum_{u\in[U]} \sum_{k\neq k^*_u} \left(\mathbb{E}\left[T_{uk}(t)\right] + \sum_{u'\neq u} \mathbb{1}\left\{k^*_{u'} = k\right\} \mathbb{E}\left[T_{u'k^*_u}(t)\right]\right) \\ &\geq \frac{U(K-1)}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)}\left((1-\alpha)\log t - \log(4KC)\right).\end{aligned}$$

Using the definition of $D(\boldsymbol{\mu})$ in the above inequality gives part (a) of the corollary.

To prove part (b), we can assume without loss of generality that a perfect matching is scheduled in every time-slot. Using this, and the fact that any server is assigned to at most one queue in every time-slot, for any $u \in [U]$, we have

$$T_{uk_u^*}(t) + \sum_{k \neq k_u^*} T_{uk}(t) = t \geq T_{uk_u^*}(t) + \sum_{u' \neq u} T_{u'k_u^*}(t),$$

which gives us

$$\sum_{k \neq k_u^*} T_{uk}(t) \geq \max \left\{ \sum_{u' \neq u} T_{uk_{u'}^*}(t), \sum_{u' \neq u} T_{u'k_u^*}(t) \right\}. \tag{31}$$

From Lemma 19 we have, for any $u' \neq u$ and for $t > \tau$,

$$\mathbb{E}\left[T_{uk_{u'}^*}(t)\right] + \mathbb{E}\left[T_{u'k_u^*}(t)\right] \geq \frac{1}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)} \left((1-\alpha)\log t - \log(4KC)\right),$$

which gives

$$\sum_{u' \neq u} \mathbb{E}\left[T_{uk_{u'}^*}(t)\right] + \mathbb{E}\left[T_{u'k_u^*}(t)\right] \geq \frac{U-1}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)} \left((1-\alpha)\log t - \log(4KC)\right).$$

Combining the above with (31), we have for $t > \tau$

$$\begin{aligned}
\sum_{k \neq k_u^*} \mathbb{E}\left[T_{uk}(t)\right] &\geq \max \left\{ \sum_{u' \neq u} \mathbb{E}\left[T_{uk_{u'}^*}(t)\right], \sum_{u' \neq u} \mathbb{E}\left[T_{u'k_u^*}(t)\right] \right\} \\
&\geq \frac{U-1}{2\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)} \left((1-\alpha)\log t - \log(4KC)\right).
\end{aligned}$$

To prove part (c), we use the fact that $\mathbb{1}\left\{k_{u'}^* = k\right\} = 0$ for any $u' \in [U]$, $K \geq k > U$. Therefore, for $t > \tau$, we have

$$\begin{aligned}
\sum_{k > U} \mathbb{E}\left[T_{uk}(t)\right] &= \sum_{k > U} \left(\mathbb{E}\left[T_{uk}(t)\right] + \sum_{u' \neq u} \mathbb{1}\left\{k_{u'}^* = k\right\} \mathbb{E}\left[T_{u'k_u^*}(t)\right] \right) \\
&\geq \frac{K-U}{\mathrm{KL}\left(\mu_{min}, \frac{\mu_{max}+1}{2}\right)} \left((1-\alpha)\log t - \log(4KC)\right),
\end{aligned}$$

which gives the required result. □

8.2.1 Late Stage: Proof of Theorem 16

The following lemma, which gives a lower bound on the queue-regret in terms of probability of sub-optimal schedule in a single time-slot, is the key result used in the proof of Theorem 16. The proof for this lemma is based on the idea that the growth in regret in a single-time slot can be lower bounded in terms of the probability of sub-optimal schedule in that time-slot.

Lemma 21. *For any problem instance characterized by $(\boldsymbol{\lambda}, \boldsymbol{\mu})$, and for any scheduling policy, and user $u \in [U]$,*

$$\Psi_u(t) \geq \lambda_u \sum_{k \neq k_u^*} \Delta_{uk} \mathbb{P}\left[\mathbb{1}\{\kappa_u(t) = k\} = 1\right].$$

Proof. For the given queueing system, consider an alternate coupled queueing system such that

1. the two systems start with the same initial condition,
2. the arrival process for both the systems is the same, and

3. the service process for the alternate system is independent of the arrival process and i.i.d. across time-slots. For each queue in the alternate system, the service offered by different servers at any time-slot could possibly be dependent on each other but has the same marginal distribution as that in the original system and is independent of the service offered to other queues.

We first show that, under any scheduling policy, the regret for the alternate system has the same distribution as that for the original system. Note that the evolution of the queues is a function of the process $(\mathbf{Z}(l))_{l\geq 1} := (\mathbf{A}(l), \boldsymbol{\kappa}(l), \mathbf{S}(l))_{l\geq 1}$. To prove that this process has the same distribution in both the systems, we use induction on the size of the finite-dimensional distribution of the process. In other words, we show that the distribution of the vector $(\mathbf{Z}(l))_{l=1}^{t}$ is the same for the two systems for all t by induction on t.

Suppose that the hypothesis is true for $t-1$. Now consider the conditional distribution of $\mathbf{Z}(t)$ given $(\mathbf{Z}(l))_{l=1}^{t-1}$. Given $(\mathbf{Z}(l))_{l=1}^{t-1}$, the distribution of $(\mathbf{A}(t), \boldsymbol{\kappa}(t))$ is identical for the two systems for any scheduling policy since the two systems have the same arrival process. Also, given $\left((\mathbf{Z}(l))_{l=1}^{t-1}, \mathbf{A}(t), \boldsymbol{\kappa}(t)\right)$, the distribution of $\mathbf{S}(t)$ depends only on the marginal distribution of the scheduled servers given by $\boldsymbol{\kappa}(t)$ which is again the same for the two systems. Therefore, $(\mathbf{Z}(l))_{l=1}^{t}$ has the same distribution in the two systems. Since the statement is true for $t=1$, it is true for all t.

Thus, to lower bound the queue-regret for any queue $u \in [U]$ in the original system, it is sufficient to lower bound the corresponding queue-regret of an alternate queueing system constructed as follows: let $\{U(t)\}_{t\geq 1}$ be i.i.d. random variables distributed uniformly in $(0,1)$. For the alternate system, let the service process for queue u and server k be given by $R_{uk}(t) = \mathbb{1}\{U(t) \leq \mu_{uk}\}$. Since $\mathbb{E}[R_{uk}(t)] = \mu_{uk}$, the marginals of the service offered by each of the servers is the same as the original system. In addition, the initial condition, the arrival process and the service process for all other queues in the alternate system are identical to those in the original system.

We now lower bound the queue-regret for queue u in the alternate system. Note that, since $\mu_u^* > \mu_{uk}\ \forall k \neq k_u^*$, we have $R_{uk_u^*}(t) \geq R_{uk}(t)\ \forall k \neq k_u^*,\ \forall t$. This implies that $Q_u^*(t) \leq Q_u(t)\ \forall t$. Now, for any given t, using the fact that $Q_u^*(t-1) \leq Q_u(t-1)$, it is easy to see that

$$Q_u(t) - Q_u^*(t) \geq \mathbb{1}\{A_u(t)=1\}\left(R_{k_u^*}(t) - \sum_{k=1}^{K}\mathbb{1}\{\kappa_u(t)=k\}R_{uk}(t)\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Q_u(t) - Q_u^*(t)] &\geq \mathbb{E}\left[\mathbb{1}\{A_u(t)=1\}\left(R_{k_u^*}(t) - \sum_{k=1}^{K}\mathbb{1}\{\kappa_u(t)=k\}R_{uk}(t)\right)\right] \\
&= \lambda_u \sum_{k\neq k_u^*} \mathbb{P}[\mathbb{1}\{\kappa_u(t)=k\}=1]\,\mathbb{P}[\mu_{uk} < U(t) \leq \mu_u^*] \\
&= \lambda_u \sum_{k\neq k_u^*} \Delta_{uk}\mathbb{P}[\mathbb{1}\{\kappa_u(t)=k\}=1].
\end{aligned}$$

□

We now use Lemma 21 in conjunction with the lower bound for the expected number of sub-optimal schedules for an α-consistent policy (Corollary 20) to prove Theorem 16.

Proof of Theorem 16. From Lemma 21 we have,

$$\begin{aligned}
\Psi_u(t) &\geq \lambda_u \sum_{k\neq k_u^*} \Delta_{uk}\mathbb{P}[\mathbb{1}\{\kappa_u(t)=k\}=1] \\
&\geq \lambda_{min}\Delta \sum_{k\neq k_u^*} \mathbb{P}[\mathbb{1}\{\kappa_u(t)=k\}=1]. \qquad (32)
\end{aligned}$$

Therefore,

$$\sum_{s=1}^{t}\sum_{u\in[U]}\Psi_u(s) \geq \lambda_{min}\Delta\sum_{u\in[U]}\sum_{k\neq k_u^*}\mathbb{E}\left[T_{uk}(t)\right].$$

We now claim that

$$\sum_{u\in[U]}\Psi_u(t) \geq \frac{U(K-1)}{8t}\lambda_{min}D(\boldsymbol{\mu})(1-\alpha) \tag{33}$$

for infinitely many t. This follows from part (a) of Corollary 20 and the following fact:

Fact 1. *For any bounded sequence $\{a_n\}$, if there exist constants C and n_0 such that $\sum_{m=1}^{n} a_m \geq C\log n\ \forall n \geq n_0$, then $a_n \geq \frac{C}{2n}$ infinitely often.*

Similarly, for any $u \in U$, it follows from parts (b) and (c) of Corollary 20 that

$$\Psi_u(t) \geq \frac{\max\{U-1, 2(K-U)\}}{8t}\lambda_{min}D(\boldsymbol{\mu})(1-\alpha) \tag{34}$$

for infinitely many t. □

8.2.2 Early Stage: Proof of Theorem 17

In order to prove Theorem 17, we first derive, in the following lemma, a lower bound on the queue-regret in terms of the expected number of sub-optimal schedules.

Lemma 22. *For any system with parameters $(\boldsymbol{\lambda}, \boldsymbol{\mu})$, any policy, and any user $u \in [U]$, the regret is lower bounded by*

$$\Psi_u(t) \geq \sum_{k\neq k^*}\Delta_{uk}\mathbb{E}\left[T_{uk}(t)\right] - \epsilon_u t.$$

Proof. Since $Q_u(0) \sim \pi_{\lambda_u,\mu_u^*}$, we have,

$$\begin{aligned}
\Psi_u(t) &= \mathbb{E}\left[Q_u(t) - Q_u^*(t)\right] \\
&= \mathbb{E}\left[Q_u(t) - Q_u(0)\right] \\
&\geq \mathbb{E}\left[\sum_{l=1}^{t} A_u(l) - S_u(l)\right] \\
&= \lambda_u t - \sum_{k=1}^{K}\mathbb{E}\left[T_{uk}(t)\right]\mu_{uk} \\
&= \lambda_u t - \left(t - \sum_{k\neq k_u^*}\mathbb{E}\left[T_{uk}(t)\right]\right)\mu *_u - \sum_{k\neq k_u^*}\mathbb{E}\left[T_{uk}(t)\right]\mu_{uk} \\
&= \sum_{k\neq k_u^*}\Delta_{uk}\mathbb{E}\left[T_{uk}(t)\right] - \epsilon_u t.
\end{aligned}$$

□

We now use this lower bound along with the lower bound on the expected number of sub-optimal schedules for α-consistent policies (Corollary 20).

Proof of Theorem 17. To prove part (a) of the theorem, we use Lemma 22 and part (a) of corollary 20 as follows: For any $\gamma > \frac{1}{1-\alpha}$, there exist constants C_4 and τ such that for all $t \in$

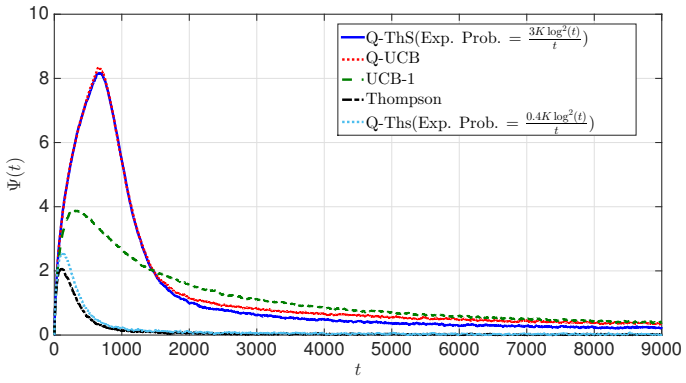


Figure 3: Comparison of queue-regret performance of Q-ThS, Q-UCB, UCB-1 and Thompson Sampling in a 5 server system with $\epsilon_u = 0.15$ and $\Delta = 0.17$. Two variants of Q-ThS are presented, with different exploration probabilities; note that $3K\log^2 t/t$ is the exploration probability suggested by theoretical analysis (which is necessarily conservative). Tuning the constant significantly improves performance of Q-ThS relative to Thompson sampling.

$[\max\{C_4 K^\gamma, \tau\}, (K-1)\frac{D(\boldsymbol{\mu})}{4\bar{\epsilon}}]$,

$$\begin{aligned}
\frac{1}{U}\sum_{u\in[U]}\Psi_u(t) &\geq \frac{\Delta}{U}\sum_{u\in[U]}\left(\sum_{k\neq k_u^*}\mathbb{E}\left[T_k(t)\right] - \epsilon_u t\right) \\
&\geq (K-1)\frac{D(\boldsymbol{\mu})}{2}\left((1-\alpha)\log t - \log(KC_4)\right) - \bar{\epsilon}t \\
&\geq (K-1)\frac{D(\boldsymbol{\mu})}{2}\frac{\log t}{\log\log t} - \bar{\epsilon}t \\
&\geq (K-1)\frac{D(\boldsymbol{\mu})}{4}\frac{\log t}{\log\log t},
\end{aligned}$$

where the last two inequalities follow since $t \geq C_4 K^\gamma$ and $t \leq (K-1)\frac{D(\boldsymbol{\mu})}{4\bar{\epsilon}}$.

Part (b) of the theorem can be similarly shown using parts (b) and (c) of corollary 20. □

Additional Discussion: As mentioned in Section 7, we note that (unstructured) Thompson sampling [19] is an intriguing candidate for future study.

In Figure 3, we benchmark the performance of Q-ThS against unstructured versions of UCB-1, Thompson Sampling and also a structured version of UCB (Q-UCB) analogous to Q-ThS. Note that there are two variants of Q-ThS displayed: the first has exploration probability $3K\log^2 t/t$, as suggested by the theory; the second has a tuned constant, with an exploration probability of $0.4K\log^2 t/t$.

It can be observed that in the early stage the unstructured algorithms perform better which is an artifact of the extra exploration required by Q-UCB and Q-ThS. In the late stage we observe that Q-UCB gives marginally better performance than UCB-1, however Thompson sampling has the best performance in both stages. This opens up additional research questions, discussed in Section 7. Q-ThS is dominated as well, but can be made to nearly match Thompson sampling by tuning the exploration probability (cf. the discussion above).

Nevertheless, it appears that Thompson sampling dominates UCB-1, Q-UCB, and the theoretically analyzed version of Q-ThS, at least over the finite time intervals considered. In some sense this is not surprising; empirically, similar observations in standard bandit problems [23, 24] are what have led to a surge of interest in Thompson sampling in the first place. Given these numerical experiments, it is important to quantify whether theoretical regret bounds can be established for Thompson sampling (e.g., in the spirit of the analysis in [25, 6, 26]).

Contextual Bandits with Latent Confounders: An NMF Approach

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

The University of Texas at Austin

July 23, 2017

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.

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, 39] that the best expected regret obtainable scales as $O\left(KL \log T\right)$.

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.

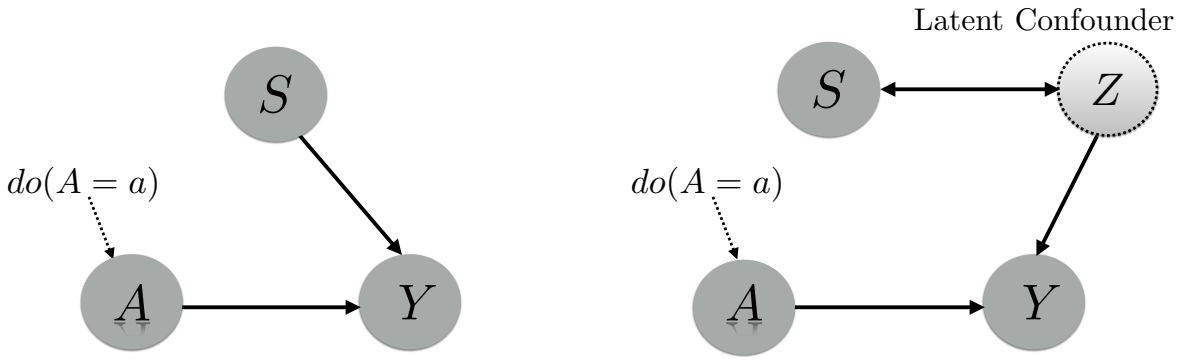
This problem of a contextual bandit with a latent structure has direct analogy with problem of designing structural *interventions* (forcing variables to take particular values) in causal graphs, a class of problems that is of increasing importance in social sciences, economics, epidemiology and computational advertising [31, 8].

A Causal Perspective: A *causal model* [31] 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* [31] 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 [35]. 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, 38] 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 $O(LK \log T)$ [39]. 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, 30, 29]. 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 [28], 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, 32] 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 Main 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 [32]. 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 $S \longleftarrow Z$ 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,

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

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* $\psi_m^1(\mathbf{P}) = \inf_{\mathbf{a} \neq 0 : \mathbf{a}^T \mathbf{1} = 0} \frac{\|\mathbf{a}^T \mathbf{P}\|_1}{\|\mathbf{a}\|_1}$.

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, 36, 35, 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*

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

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 $\Pr_{|S|=m'}(\sigma_{\min}(\mathbf{P}_{S,:}) \geq \rho) \geq 1 - \epsilon$ 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 $\Pr_{|S|=m'}(\psi_m^1(\mathbf{P}_{:,S}) \geq \rho) \geq 1-\epsilon$ 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 [32]. 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,*

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

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, 36, 35, 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 [32]. 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 some 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 [32] 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 [33] 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* $k \neq k^*(s)$ *and* $s \in \mathcal{S}$, *where* $\alpha \in (0,1)$ *and* $T(s) = \sum_{t=1}^{T} \mathbb{1}\{S_t = s\}$

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* $D(\mathbf{U}) = O\left(1/\Delta\right)$ *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 as follows.

(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 [32, 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 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.

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 Simulations

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:

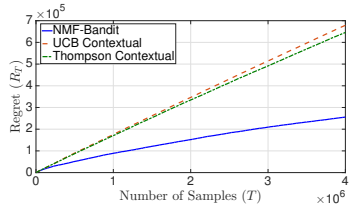
S1 : The arrival of the contexts *cannot* be controlled by the algorithm and the regret is w.r.t the best arm which is context *dependent*. This is strongly motivated by the causal setting discussed with real world scenarios in Section **??**.

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.

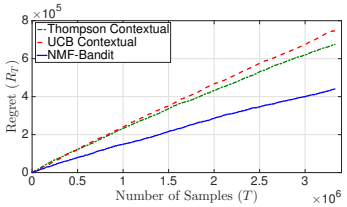
Synthetic Data-Sets : In order to generate the synthetic reward matrix $\mathbf{U}$, the parameter L, K, m are chosen. The $L \times m$ matrix $\mathbf{A}$ is then generated by picking each row uniformly at random from the m-dimensional simplex. The $m \times K$ matrix $\mathbf{W}$ is generated with 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.

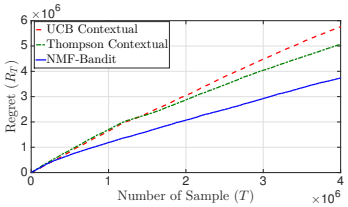
Real World Data-Sets : We use the Movielens 1M [21] and the Book Crossing [40] data-sets for our real world experiments. A subset of dimension 2000×2000 is chosen from



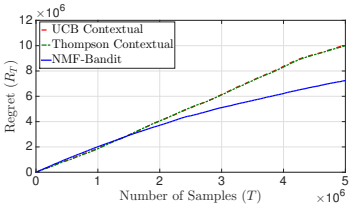
(a) Synthetic data-set with $L = 455$, $K = 210$ and $m = 7$. The rewards are Uniform around the means with a support of length 0.4. Setting : **S1**



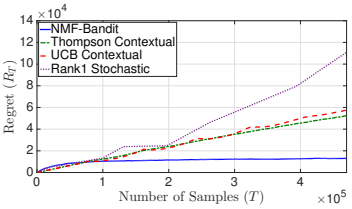
(b) Synthetic data-set with $L = 300$, $K = 145$ and $m = 3$; the rewards are Bernoulli with the given means. Setting : **S1**



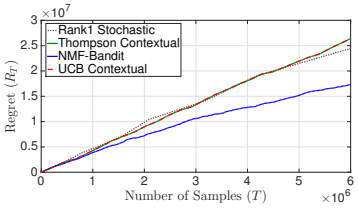
(c) A random subset of Movielens data-set with $L = 1600$, $K = 750$. The rewards are Uniform around the means with a support of length 2. In our algorithm we set $m = 10, m' = 20$ and $\theta = 3$. Setting : **S1**



(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. In our algorithm 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.

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 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 has been specified 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.

We validate our algorithm on real and synthetic datasets and show superior performance with respect to the baselines considered. 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. This is a natural setting and we expect that it will lead to an interesting research direction.

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

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

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

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

[32] 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.

[33] 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.

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

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

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

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

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

[39] 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.

[40] 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.

A Appendix

A.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 represent the sampled entries from different subsets of arms in a structured manner.

A.1.1 Matrix Sampling

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

- Given that the context is s, sample a uniform random variable κ with support $\{a_1 \dots a_p\}$, 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 $\hat{\mathbf{u}}[s]_k = Y_\kappa \mathbf{e}_\kappa$.

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

A.1.2 Arms to be sampled during *explore*

Before we present the pseudocode, we define the sampling matrices $\{\mathbf{G}(0), \mathbf{G}(1), \cdots, \mathbf{G}(l+1)\}$. Recall that any subset of arms can be encoded in a sampling matrix. $\mathbf{G}(0)$ 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*. $\mathbf{G}(i)$ corresponds to the subset $R(s_t)$ is $s_t \in S(i)$. Let $l = \lfloor K/m \rfloor$ and $r = K \bmod (m)$. 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 $\{S(1), S(2), ..., S(l+1)\}$ of size $2m'$ each. The elements of the set $S(j)$ will be denoted as $S(j) = \{s_1(j), s_2(j) \cdots, s_{2m'}(j)\}$. In Step 2 of *explore*, if $s_t \in S(i)$, then $R(s_t) = \{(i-1)m, (i-1)m+1, \cdots \max(im-1, K)\}$. 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.

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

2. $\mathbf{G}(i)$: An $K \times m$ matrix such that,

$$G(i)_{kj} = \left\{ \begin{array}{ll} 1, & \text{if } k = (i-1)m + j \text{ for } j \in \{1, \cdots, m\} \\ 0, & \text{otherwise} \end{array} \right\}$$

when $i \in \{1, 2, \cdots, l\}$.

3. $\mathbf{G}(l+1)$: An $K \times r$ matrix defined as follows:

$$G(l+1)_{kj} = \left\{ \begin{array}{ll} 1, & \text{if } k = (lm + j) \text{ for } j \in \{1, \cdots r\} \\ 0, & \text{otherwise} \end{array} \right\}$$

In words, $\mathbf{G}(i)$ 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.

A.1.3 Representation of the collected Samples

In what follows, let the mean of samples collected through $\mathbf{G}(0)$ till time t be collected in a $L \times 2m'$ matrix $\hat{\mathbf{F}}'(t)$ such that $\mathbb{E}\left[\hat{\mathbf{F}}'(t)\right] = (1/2m')\mathbf{F} = (1/2m')\mathbf{U}\mathbf{G}(0)$ as detailed in Section A.1.1. Let $\hat{\mathbf{F}}(t) = 2m'\hat{\mathbf{F}}'(t)$. Let the samples collected from $\mathbf{G}(i)$ be stored in a $2m' \times m$ matrix $\hat{\mathbf{M}'}_i(t)$ such that $\mathbb{E}\left[\hat{\mathbf{M}'}_i(t)\right] = \frac{1}{m}\mathbf{A}_{S(i),:}\mathbf{W}\mathbf{G}(i)$ for all $i \in \{1, 2, ..., l+1\}$. Let $\hat{\mathbf{M}}_i(t) = m\hat{\mathbf{M}'}_i(t)$ be the scaled version.

A.1.4 Pseudocode

We present a detailed pseudo-code of our algorithm as Algorithm 1. For the sake of completeness we include the robust version of the Hottopix algorithm [18] which is used as a sub-program in Algorithm 1. The following LP is fundamental to the Hottopix algorithm,

$$\begin{aligned} &\min_{\mathbf{C} \in \mathbb{R}_+^{f \times n}} \mathbf{p}^T \text{diag}(\mathbf{C}) \\ &\text{s.t. } \left\| \tilde{\mathbf{X}} - \mathbf{C}\tilde{\mathbf{X}} \right\|_{\infty,1} \leq 2\epsilon \\ &\text{and } C_{ii} \leq 1,\ C_{ji} \leq C_{ii}\ \forall i, j \in [L] \end{aligned} \tag{3}$$

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

Algorithm 1 NMF-Bandit - An ϵ-greedy algorithm for Latent Contextual Bandits

At time t,
Observe context $S_t = s_t$
Let $\mathsf{E}(t) \sim \text{Ber}(\epsilon_t)$
if $\mathsf{E}(t) = 1$ **then**
    Explore: Let $H_t \sim \left\{ \begin{array}{ll} \text{Ber}\left(\frac{2m'}{r+2m'}\right), & \text{if } s_t \in S(l+1) \\ \text{Ber}\left(\frac{2m'}{m+2m'}\right), & \text{otherwise} \end{array} \right\}$.
    If $H_t = 1$ sample an arm according to the matrix sampling technique applied to matrix $\mathbf{G}(0)$ and update $\hat{\mathbf{F}}(t)$.
    If $H_t = 0$ sample an arm according to the matrix sampling technique applied to matrix $\mathbf{G}(i)$ if $s_t \in S(i)$ for $i \in \{1, 2, \cdots, l+1\}$ and update $\hat{\mathbf{M}}_i(t)$. If s_t is not in any of these sets then choose an arm at random.
else
    Exploit:
    Let us compute,

$$\hat{\mathbf{W}}(t) = \text{Hottopix}(\mathbf{F}(t), m, 2m'\gamma(t)).$$
$$\hat{\mathbf{A}}(t) = \underset{\mathbf{Z}\geq\mathbf{0}, \text{rowsum}(\mathbf{Z})=\mathbf{1}}{\text{argmin}} \left\| \mathbf{F}(t) - \mathbf{Z}\hat{\mathbf{W}}(t) \right\|_{\infty,1}.$$

    Let $\hat{\mathbf{W}}(t) \in \mathbb{R}^{m\times K}$ be such that,

$$\hat{\mathbf{W}}(t)_{:,(i-1)m:im-1} = \underset{\mathbf{X}_{m\times m}}{\text{argmin}} \left\| \hat{\mathbf{A}}(t)_{S(i),:}\mathbf{X} - \hat{\mathbf{M}}_i(t) \right\|_2, \ \forall i \in \{1, 2, .., l\}$$
$$\hat{\mathbf{W}}(t)_{:,lm:K} = \underset{\mathbf{X}_{m\times r+1}}{\text{argmin}} \left\| \hat{\mathbf{A}}(t)_{S(l),:}\mathbf{X} - \hat{\mathbf{M}}_{l+1}(t) \right\|_2$$

    Compute $\hat{\mathbf{U}}(t) = \hat{\mathbf{A}}(t)\hat{\mathbf{W}}(t)$. Play the arm a_t such that,

$$a_t = \arg\max_a \hat{\mathbf{U}}(t)_{s_t,a}$$

end if

Algorithm 2 Hottopix($\tilde{\mathbf{X}}, m, \epsilon$)

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

A.2 Theoretical Insights

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 l_∞-norm guarantees on the recovered matrix [20]. In order to ensure such noise guarantees, we require a very large number of samples in order for estimates to concentrate. 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 ℓ_1-WStRIP property of $\mathbf{W}$, we can ensure that NMF on these samples (which are basically a noisy version of $\mathbf{U}_{:,S}$) gives us a good estimate of $\mathbf{A}$ at time t; this estimate is denoted by $\hat{\mathbf{A}}(t)$. 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 $\mathbf{U} = \mathbf{AW}$. Therefore, an initial approach would be to use the current estimate of $\mathbf{A}$ along with samples of the rewards, and directly recover $\mathbf{W}$. This however will not work due to lack of concentrations. First, the estimate of $\mathbf{A}$ 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 $\mathbf{U}$ 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 $\mathbf{W}$. 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 ℓ_2-WStRIP property of $\mathbf{A}$ we can get a good estimate of $\mathbf{W}$, denoted by $\hat{\mathbf{W}}(t)$ 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 l_∞-norm bounds on the errors in $\hat{\mathbf{A}}(t)$ and $\hat{\mathbf{W}}(t)$, imply that $\left\|\hat{\mathbf{U}}(t)-\mathbf{U}\right\|_{\infty,\infty} < \Delta/2$ with probability at least $1-O(\frac{Lm'}{t})$ provided ϵ_t is sufficiently big (see proof of Theorem 8). Here $\Delta = \min_{s\in[L]}(u^*(s)-\max_{k\neq k^*(s)} U_{s,k})$. This essentially implies that the correct arm is pulled at time t w.h.p if the algorithm decides to *exploit*.

A.3 Description of Generative Models for matrices W and A

The model for $\mathbf{W}$ and $\mathbf{A}$ are both very similar with deterministic and random parts. The technical description of the model given below is complex due to the following two reasons:

1. **Fact 1:** Rows of $\mathbf{A}$ must sum to 1.

2. **Fact 2:** The rows of $\mathbf{W}$ shifted by an arbitrary vector $\mathbf{m} \in \mathbb{R}^{1\times K}$ does not affect the NMF algorithms employed. The setting is invariant to such a shift.

1. *Random+Deterministic Composition*:

   (a) We assume that columns $\mathbf{W}_{:,D}$ corresponding to the column index set $D \subseteq [K]$, $|D| \leq K/(32m)$ is arbitrary and deterministic. $0 \leq W_{i,j} \leq 1,\ j \in D$. The maximum entry in every row of $\mathbf{W}$ is assumed to be contained in the deterministic part.

   (b) Similarly, $\mathbf{A}_{E,:}$ where $E \subseteq [L]$ is arbitrary and deterministic. Let $|E| \leq \rho L$. $\rho = 1/18$. Row sum of every row of $\mathbf{A}_{E,:}$ is 1. In order to ensure separability [32] we assume that there is a subset $M \subseteq E : |M| = m$ such that $\mathbf{A}_{M,:} = \mathbf{I}_{m\times m}$. For all $i \in E - M$, $0 \leq A_{ij} \leq \gamma < 1$.

2. *Bounded randomness in the random part:*

$$\mathbf{W}_{:,D^c} = \mathbf{1} * \mathbf{m}^T + \mathbf{R}_{:,D^c} + \tilde{\mathbf{W}}_{:,D^c} \tag{4}$$

   (a) (i,j)-th entry of $\tilde{\mathbf{W}}_{:,D^c}$ is an independent mean zero sub-gaussian entry with variance q, and bounded support and sub-gaussian parameter $c(q)$. $\mathbf{m} \in \mathbb{R}^{|D^c|\times 1}$ is an arbitrary deterministic vector [1] .

   (b) $\mathbf{R}_{:,D^c}$ is a deterministic perturbation matrix satisfying $\|\mathbf{R}_{:,j}\|_2 \leq \frac{1}{5},\ \forall j \in D^c$. The support parameters for $\tilde{\mathbf{W}}_{:,D^c}$, $\mathbf{m}$ and $\mathbf{R}_{:,D^c}$ are chosen such that $0 \leq W_{i,j} \leq 1$ a.s. , $\forall j \in D^c$

$\mathbf{A}_{E^c,:}$ is a matrix which is a row-normalized version of another random matrix $\tilde{\mathbf{A}}$. We first describe the random model on the $|E^c| \times m$ matrix $\tilde{\mathbf{A}}$. Like in the case for model of $\mathbf{W}$,

$$\tilde{\mathbf{A}} = \mathbf{N} + \hat{\mathbf{A}} \tag{5}$$

   (a) $\hat{\mathbf{A}}$ is a matrix with independent mean zero sub-gaussian entries each with variance q, and bounded support and sub-gaussian parameter $c(q)$.

   (b) We denote the matrix of means by $\mathbf{N}$ consisting of the parameters n_{ij}. The ℓ_2 norm of every row of $\mathbf{N}$ is at most $\frac{1}{5}$. The support, sub-gaussian parameter and the matrix of means $\mathbf{N}$ are chosen such that $1/m \leq \tilde{A}_{ij} \leq \gamma < 1$ a.s. The stricter condition (in the lower bound) ensures that after normalization by the row sum, $A_{ij} \leq \gamma < 1,\ i \in E^c$.

A.4 Projection onto a Low Dimensional Space

In this section, we will prove some properties of the matrix $\mathbf{F} = \mathbf{UG}(0) = \mathbf{AWG}(0)$ where $\mathbf{G}(0)$ is a $K \times 2m'$ as defined in Section A.1.1. From the definition in Section 2.1, $\mathbf{A}$ contains a $\mathbf{I}_{m\times m}$ sub-matrix corresponding to the rows in $\mathcal{Z}$. Further, the row sum of every row of $\mathbf{A}$

[1] This is introduced to respect Fact 2 in Section A.3

is 1. This means that the rows of $\mathbf{U}$ consists of points in the convex hull of extreme points, i.e. the rows of $\mathbf{W}$, together with the extreme points themselves.

The extreme points in $\mathbf{W}$ are mapped to extreme points in $\mathbf{WG}(0)$. We also show that the new set of extreme points $\mathbf{WG}(0)$ also satisfy what is called the simplical property when $\mathbf{W}$ satisfies the assumptions in Section A.3.

When the entries in $\mathbf{W}$ are random and independent bounded random variables as in Section 2.4, we show that ℓ_1 distance of any non-zero vector $\mathbf{a}$ such that $\mathbf{a}^T\mathbf{1} = 0$ is preserved under the map $\mathbf{a}^T\mathbf{WG}(0)$with high probability over $\mathbf{W}$ for any fixed $\mathbf{G}(0)$. We need some results relating to sub-gaussianity of the matrix $\mathbf{W}$ which we deal with in the next subsection.

A.5 Sub-gaussianity of a matrix with bounded i.i.d random entries

Definition 7. *[16] 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 8. *[16] A random vector $\mathbf{Y} \in \mathbb{R}^n$ is isotropic if $\mathbb{E}[(\mathbf{Y}^T\mathbf{x})^2] = \mathbb{E}[\mathbf{x}^T\mathbf{x}],\ \forall \mathbf{x} \in \mathbb{R}^n$. It is sub-gaussian with parameter c if the scalar random variable $\mathbf{Y}^T\mathbf{x}$ is sub-gaussian with parameter c for all $\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1$, i.e. $\mathbb{E}[\exp(t(\mathbf{Y}^T\mathbf{x}))] \leq \exp(-ct^2), \forall t \in \mathbb{R},\ \forall \|\mathbf{x}\|_2 = 1$.*

Lemma 1. *[16],[34] 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 $\mathbf{Y} \in \mathbb{R}^n$ where each entry is drawn i.i.d from a mean zero, unit variance and a sub-gaussian distribution with parameter c. Then $\mathbf{Y}$ is a sub-gaussian isotropic vector with the same sub-gaussian parameter c*

Remark: The first part is from Theorem 9.9 in [34] while the second part is from Lemma 9.7 from [16].

Lemma 2. *[16] Let $\mathbf{P}$ and $\mathbf{Q}$ be two matrices of the same dimensions. Let $\sigma_{\min}$ and $\sigma_{\max}$ be the largest and smallest singular values of a matrix respectively. Then,*

$$|\sigma_{\min}(\mathbf{P}) - \sigma_{\min}(\mathbf{Q})| \leq \sigma_{\max}(\mathbf{P} - \mathbf{Q}) \tag{6}$$

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

$$\sigma_{\max}(\mathbf{P}^T\mathbf{P} - \mathbf{I}_{n\times n}) \leq \delta \Rightarrow \sigma_{\min}(P) \geq \sqrt{(1-\delta)} \tag{7}$$

Lemma 3. *[16] Consider an $m \times s$ matrix $\mathbf{P}$ with every row being a random independent sub-gaussian isotropic vector with sub-gaussian parameter c. Let $m > s$, then:*

$$\Pr\left(\sigma_{\max}\left(\frac{1}{m}\mathbf{P}^T\mathbf{P} - \mathbf{I}_{s\times s}\right) \geq \delta\right) \leq 2\exp(-\frac{3\tilde{c}}{4}\delta^2 m + \frac{7s}{2}) \tag{8}$$

Further,

$$\Pr\left(\sigma_s(\mathbf{P}) \leq \sqrt{m}\sqrt{(1-\delta)}\right) \tag{9}$$

$$\leq \Pr\left(\sigma_{\max}\left(\frac{1}{m}\mathbf{P}^T\mathbf{P} - \mathbf{I}_{s\times s}\right) \geq \delta\right)$$

$$\leq 2\exp(-\frac{3\tilde{c}}{2}\delta^2 m + \frac{7s}{2}) \tag{10}$$

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 [16] and also from combining Lemma 9.8 and Lemma 9.9 in [16]. The second follows from applying Lemma 2

Definition 9 ([32]). *Let us consider a matrix* $\mathbf{M}$ *which is* $p \times q$ *where* $p \leq q$. *Let* $\mathbf{m}_i \in \mathbb{R}^{1\times p}$ *be the* i*-th row of the matrix* $\mathbf{M}$. *The matrix* M *is* α*-simplical if* $\min_{i\in\{1\cdots p\}} \min_{\mathbf{x}\in \text{conv}(\{\mathbf{m}_1\cdots\mathbf{m}_p\}\setminus\mathbf{m}_i)} \|\mathbf{m}_i - \mathbf{x}\|_1 \geq \alpha$. *In other words, every row is at least* α *far away in* ℓ_1 *distance from the convex hull of other points.*

A.6 Results regarding sub-matrices of W

The following results hold for $\mathbf{WG}(0))$ since $\mathbf{WG}(0) = \mathbf{W}_{:,S}$ when $S = \{a_1 \dots a_{m'}\}$ is the set of column indices associated with $\mathbf{G}(0)$ as in Section A.1.

Theorem 5. *Let* $\mathbf{W}$ *follow the random generative model in Section 2.4. Let* $S \subseteq D^c$. *Let* $|S| = m' \geq \frac{512}{21\tilde{c}} m \log(eK)$,

$$\psi_m(\mathbf{W}_{:,S}) \geq \left(\frac{11}{20}\right)\sqrt{m'} \tag{11}$$

with probability at least $1 - \frac{2}{K^{7m/2}}$ *over the randomness in* $\mathbf{W}$. *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.*

Proof. According to the random generative model for $\mathbf{W}$ in Section 2.4, $\mathbf{W}_S = \tilde{\mathbf{W}}_{:,S} + \mathbf{1}\mathbf{m}_S^T + \mathbf{R}_S$. Here, $\tilde{\mathbf{W}}_{:,S}$ 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 $\mathbf{m}_S$ refer to the vector restricted to co-ordinate in S. Applying Lemma 3 to the sub-gaussian matrix $(m' \times m)$ $\tilde{\mathbf{W}}_{:,\mathbf{S}}$ with $m' \geq \frac{512}{21\tilde{c}} m \log(eK)$ and setting $\delta = 7/16$, we have:

$$\begin{aligned}\Pr\left(\sigma_m(\tilde{\mathbf{W}}^T_{:,\mathbf{S}}) \leq \frac{3}{4}\sqrt{m'}\right) &\leq 2\exp(-\frac{7}{2} m \log(K)) \\ &\leq 2K^{-7m/2}.\end{aligned}$$

Now, applying Lemma 2, we have:

$$\begin{aligned}|\sigma_m(\mathbf{R}_{:,S} + \tilde{\mathbf{W}}_{:,S}) - \sigma_m(\tilde{\mathbf{W}}_{:,S})| &\leq \sigma_{max}(\mathbf{R}_{:,S}) \\ &\leq \|\mathbf{R}_{:,S}\|_F \\ &\leq \frac{1}{5}\sqrt{m'}\end{aligned}$$

Combining the above two equations, we have:

$$\begin{aligned}&\Pr\left(\sigma_m(\tilde{\mathbf{W}}_{:,S} + \mathbf{R}_{:,S}) \leq \left(\frac{3}{4} - \frac{1}{5}\right)\sqrt{m'}\right) \\ &\leq 2\exp(-\frac{7}{2} m \log(K)) \leq 2K^{-7m/2}.\end{aligned}$$

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

$$\begin{aligned}
&\inf_{\mathbf{a}\neq 0:\mathbf{a}^T\mathbf{1}=0} \frac{\|\mathbf{a}^T\mathbf{W}_{:,S}\|_2}{\|\mathbf{a}\|_2} && (12)\\
&= \inf_{\mathbf{a}\neq 0:\mathbf{a}^T\mathbf{1}=0} \frac{\|\mathbf{a}^T(\mathbf{1}\mathbf{m}_S^T + \tilde{\mathbf{W}}_{:,S} + \mathbf{R}_{:,S})\|_2}{\|\mathbf{a}\|_2} \\
&\underset{(\mathbf{a}^T\mathbf{1}=0)}{=} \inf_{\mathbf{a}\neq 0::\mathbf{a}^T\mathbf{1}=0} \frac{\|\mathbf{a}^T(\tilde{\mathbf{W}}_{:,S} + \mathbf{R}_{:,S})\|_2}{\|\mathbf{a}\|_2} \\
&\geq \sigma_m(\mathbf{R}_{:,S} + \tilde{\mathbf{W}}_{:,S}) && (13)
\end{aligned}$$

□

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

$$\psi_m^1(\mathbf{W}_{:,S}) = \inf_{\mathbf{a}\neq 0:\mathbf{a}^T\mathbf{1}=0} \frac{\|\mathbf{a}^T\mathbf{W}_{:,S}\|_1}{\|\mathbf{a}\|_1} \geq \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}} \tag{14}$$

with probability at least $1 - 2K^{-7m/2}$ *over the randomness in* $\mathbf{W}$. *Further, rows of* $\mathbf{W}_{:,S}$ *is* $\psi_m^1(\mathbf{W}_{:,S})$*-simplical*

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

$$\begin{aligned}
\|\mathbf{a}^T(\mathbf{W}_{:,S})\|_1 &\geq \|\mathbf{a}^T[\mathbf{W}_{:,S_1}\mathbf{W}_{:,S_2}]\|_2 && (15)\\
&\underset{(a)}{\geq} \|\mathbf{a}^T\mathbf{W}_{:,S_2}\|_2 - \|\mathbf{a}^T\mathbf{W}_{:,S_1}\|_2 \\
&\underset{(b)}{\geq} \|\mathbf{a}\|_2\psi_m(\mathbf{W}_{S_2}) - \|\mathbf{a}\|_2\sqrt{m\frac{2m'}{16m}} && (16)\\
&\underset{(c)}{\geq} \|\mathbf{a}\|_1\frac{\sqrt{15m'}}{\sqrt{8m}}\left(\frac{3}{4} - \frac{1}{5} - \frac{1}{\sqrt{15}}\right) \\
&\geq \left(\frac{3}{4} - \frac{8}{15}\right)\frac{\sqrt{15m'}}{\sqrt{8m}} \text{ w.p. } 1 - \frac{2}{K^{7m/2}} && (17)
\end{aligned}$$

Justifications of the above chain are: (a)- Triangle inequality for the norm $\|\cdot\|_2$. (b)- Definition of $\psi_m(\cdot)$ and $\|\mathbf{a}^T\mathbf{W}_{S_1}\|_2 \leq \|\mathbf{a}^T\|_2|\mathbf{W}_{S_1}\|_F \leq \sqrt{m|S_1|}\|\mathbf{a}^T\|_2$. (c)- $\|\cdot\|_2 \geq \frac{\|\cdot\|_1}{\sqrt{m}}$ and applying Theorem 5 because $S_2 \subseteq D^c$ and $|S_2| \geq \frac{512}{21\tilde{c}} m\log(eK)$.

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

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

From the definition for an α-simplical matrix (Definition 9), it is enough to show that for any $\mathbf{r}^{-i}$, $\|\mathbf{r}^{-i}\mathbf{W}_S\|_1 \geq \psi_m^1(\mathbf{W}_{:,S})$. We prove this as follows:

$$\|\mathbf{r}^{-i}\mathbf{W}_{:,S}\|_1 \underset{(\|\mathbf{r}^{-i}\|_1\geq 1)}{\geq} \psi_m^1(\mathbf{W}_{:,S}) \tag{19}$$

□

A.6.1 Choosing a good S for $\mathbf{G}(0)$

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 $m' = \frac{512}{21\tilde{c}} m \log(eK)$. Then, $\Pr\left(|S \bigcap D| \leq \frac{2m'}{16m}\right) \leq \exp(-c_1 \log(eK))$ for constant $c_1 > 0$ that depends on $\tilde{c}$.*

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 $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = \frac{|D|}{K} = \mu \geq \frac{1}{32m}$. 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 [23]. In particular, the following inequality derived from moment generating functions holds [24] for any $\delta > 0$:

$$\begin{aligned}
&\Pr\left(\sum X_i \geq (1+\delta)2m'\mu\right) \\
&\leq \Pr\left(\sum X_i \geq (1+\delta)2m'\mu\right) \\
&\leq \exp\left(\delta 2m'\mu\right)\left(1+\delta\right)^{-(1+\delta)2m'\mu}.
\end{aligned}$$

Let us take $\delta = 1$. Therefore, $\Pr\left(|S \bigcap D| \geq \frac{2m'}{16m}\right) \leq \left(\frac{4}{e}\right)^{-\frac{32}{21\tilde{c}}\log(eK)} \leq \frac{1}{(eK)^{\frac{32}{21\tilde{c}}\log(4/e)}}$.

□

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

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}}\left[\mathbb{P}_{\mathbf{S}}\left(\psi_m^1(\mathbf{W}_{:,S}) < \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}}\right)\right] \\
&\leq \exp(-c_1\log(eK)) + 2K^{-7m/2} \\
&\leq 2\exp(-c_1\log(eK)
\end{aligned}$$

Now by Markov's inequality this implies that,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{W}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\psi_m^1(\mathbf{W}_{:,S}) < \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}}\right) \geq 2\exp(-\frac{c_1}{2}\log(eK))\right)\right] \\
&\leq \frac{\exp(-c_1\log(eK)}{\exp(-\frac{c_1}{2}\log(eK))} \\
&\leq \exp(-\frac{c_1}{2}\log(eK))
\end{aligned}$$

This implies the following chain:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{W}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\psi_m^1(\mathbf{W}_{:,S}) > \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}}\right) \leq 1-2\exp(-\frac{c_1}{2}\log(eK))\right)\right] \\
&\leq \exp(-\frac{c_1}{2}\log(eK)) \\
&\Rightarrow \mathbb{P}_{\mathbf{W}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\psi_m^1(\mathbf{W}_{:,S}) > \left(\frac{13}{60}\right)\frac{\sqrt{15m'}}{\sqrt{8m}}\right) \geq 1-2\exp(-\frac{c_1}{2}\log(eK))\right)\right] \\
&\geq 1-\exp(-\frac{c_1}{2}\log(eK))
\end{aligned}$$

This proves that with probability at least $1-\exp(-\frac{c_1}{2}\log(eK))$ the ℓ_1-WStRIP condition is satisfied with the said parameters. □

A.7 Results regarding sub-matrices of A

We assume that $\mathbf{A}$ 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 ($|S| = 2m'$) which is a mix of random and the deterministic columns. The proofs follow closely those of $\mathbf{W}$ in the previous section.

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

$$\sigma_m\left(\mathbf{A}_{S,:}\right) \geq \frac{\sqrt{m'}}{m}\left(\frac{1}{20}\right) \text{ w.p } 1-\frac{2}{L^{7m/2}}. \tag{20}$$

Proof. Let $\tilde{S}_2$ be the set of rows in the random matrix $\tilde{\mathbf{A}}$ that corresponds to the rows S_2 in $\mathbf{A}$. Here, $\hat{\mathbf{A}}_{\tilde{S}_2,:}$ 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 $(|\tilde{S}_2| \times m)$ $\hat{\mathbf{A}}_{\tilde{\mathbf{S}}_2,:}$ with $|\tilde{S}_2| \geq m' \geq \frac{512}{21\tilde{c}} m\log(eL)$ and setting $\delta = 7/16$, we have:

$$\begin{aligned}
\Pr\left(\sigma_m(\hat{\mathbf{A}}_{\tilde{\mathbf{S}}_2,:}) \leq \frac{3}{4}\sqrt{m'}\right) &\leq 2\exp\left(-\frac{7m\log(L)}{2}\right) \\
&\leq 2L^{-7m/2}.
\end{aligned}$$

Now, consider the following matrix: $[\frac{1}{m}\left(\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\right)\mathbf{A}_{S_1,:}]$. First, note that according to the model in Section 2.4, rows of $\mathbf{A}_{S_1,:}$ sum to 1. Therefore, we have the following chain

for any non zero vector $\mathbf{a} \in \mathbb{R}^{1\times m}$:

$$\begin{aligned}
&\|[\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\mathbf{A}_{S_1,:}]\mathbf{a}\|_2 \geq \|\mathbf{N}_{\tilde{S}_2,:} + \frac{1}{m}\hat{\mathbf{A}}_{\tilde{S}_2,:}\mathbf{a}\|_2 \qquad (21)\\
&- \|\mathbf{A}_{S_1,:}\mathbf{a}\|_2\\
&\geq \|\left(\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\right)\mathbf{a}\|_2 - \sqrt{\sum_{i\in S_1}\|\mathbf{A}_{i,:}\|_2^2\|\mathbf{a}\|_2^2}\\
&\geq \|\left(\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\right)\mathbf{a}\|_2 - \sqrt{\sum_{i\in S_1}\|\mathbf{A}_{i,:}\|_1^2\|\mathbf{a}\|_2^2}\\
&\geq \|\left(\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\right)\mathbf{a}\|_2 - \sqrt{2\rho m'}\|\mathbf{a}\|_2\\
&\geq \|\hat{\mathbf{A}}_{\tilde{S}_2,:}\mathbf{a}\|_2 - \|\mathbf{N}_{\tilde{S}_2,:}\mathbf{a}\|_2 - \sqrt{2\rho m'}\|\mathbf{a}\|_2\\
&\geq \sigma_m\left(\hat{\mathbf{A}}_{\tilde{S}_2,:} - \sqrt{2m'(1-\frac{1}{9})}\frac{1}{5} - \sqrt{2\frac{1}{9}m'}\right)\|\mathbf{a}\|_2\\
&\geq \left(\frac{3}{4} - \frac{1}{5} - \frac{\sqrt{2\frac{1}{9}}}{\sqrt{1-\frac{1}{9}}}\right)\sqrt{2m'(1-\frac{1}{9})} \text{ w.p. } 1-2L^{-7m/2}.\\
&\geq \left(\frac{3}{4} - \frac{1}{5} - \frac{1}{2}\right)\sqrt{m'} \text{ w.p. } 1-2L^{-7m/2}. \qquad (22)
\end{aligned}$$

Now, we normalize the every row of $[\mathbf{N}_{\tilde{S}_2,:} + \hat{\mathbf{A}}_{\tilde{S}_2,:}\mathbf{A}_{S_1,:}]$ to get $[\mathbf{A}_{S_2,:}\mathbf{A}_{S_1,:}] = \mathbf{A}_S\mathbf{P}$ where $\mathbf{P}$ 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,

$$\sigma_m\left(\mathbf{A}_S\right) = \sigma_m\left(\mathbf{A}_S\mathbf{P}\right) \geq \frac{\sqrt{m'}}{m}\left(\frac{3}{4} - \frac{1}{5} - \frac{1}{2}\right) \text{ w.p } 1-2L^{-7m/2}.$$

□

A.7.1 Choosing a good $S(i)$ for a $\mathbf{G}(i)$

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 $m' = \frac{512}{21\tilde{c}}m\log(eL)$. Then, $\Pr\left(|S\bigcap E| \leq \frac{2m'}{9}\right) \leq \exp(-c_2 m\log(eL))$ for constant $c_2 > 0$ that depends on $\tilde{c}$.*

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

$$\Pr\left(|S\bigcap E| \geq \frac{2m'}{9}\right) \leq \frac{1}{(eL)^{\frac{512\log(4/e)}{189\tilde{c}}m}}. \qquad (23)$$

□

Proof of Theorem 3. From Theorem 7 and Lemma 5 we have,

$$\mathbb{E}_{\mathbf{A}}\left[\mathbb{P}_{\mathbf{S}}\left(\sigma_m(\mathbf{A}_{S,:}) < \frac{\sqrt{m'}}{m}\left(\frac{1}{20}\right)\right)\right] \leq 3\exp(-c_2' m\log(eL))$$

Now by Markov's inequality this implies that,

$$\begin{aligned}&\mathbb{P}_{\mathbf{A}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\sigma_m(\mathbf{A}_{S,:}) < \frac{\sqrt{m'}}{m}\left(\frac{1}{20}\right)\right) \geq \exp(-\frac{c_2'}{2}m\log(eL))\right)\right]\\ &\leq 3\frac{\exp(-c_2' m\log(eL))}{\exp(-\frac{c_2'}{2}m\log(eL))}\\ &\leq 3\exp(-\frac{c_2'}{2}m\log(eL))\end{aligned}$$

This implies the following chain:

$$\begin{aligned}&\mathbb{P}_{\mathbf{A}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\sigma_m(\mathbf{A}_{S,:}) > \frac{\sqrt{m'}}{m}\left(\frac{1}{20}\right)\right) \leq 1-\exp(-\frac{c_2'}{2}m\log(eL))\right)\right]\\ &\leq 3\exp(-\frac{c_2'}{2}m\log(eL))\\ &\Rightarrow \mathbb{P}_{\mathbf{A}}\left[\left(\mathbb{P}_{\mathbf{S}}\left(\sigma_m(\mathbf{A}_{S,:}) > \frac{\sqrt{m'}}{m}\left(\frac{1}{20}\right)\right) \geq 1-\exp(-\frac{c_2'}{2}m\log(eL))\right)\right]\\ &\geq 1-3\exp(-\frac{c_2'}{2}m\log(eL))\end{aligned}$$

This proves that with probability at least $1-\exp(-\frac{c_2'}{2}m\log(eL))$ the ℓ_2-WStRIP condition is satisfied with the said parameters. □

A.8 Noisy NMF in Low dimensions

In this section we enhance the guarantees of the robust Hottopix algorithm from [18] provided $\mathbf{W}$ satisfies ℓ_1-WStRIP and the subset S chosen by Algorithm 1 is good as in Section 4.

Lemma 6. *Suppose* $\mathbf{W}$ *satisfies* ℓ_1*-WStRIP with parameter* $(\delta, \rho_1, 2m')$ *and the subset* S *of its columns* ($|S| = 2m'$) *satisfies* $\psi_m^1(\mathbf{W}_{S,:}) \geq \rho_1$*. Consider a matrix* $\tilde{\mathbf{X}} = \mathbf{A}\mathbf{W}_{:,S} + \mathbf{N}$ *such that* $\|\mathbf{N}\|_{\infty,1} \leq \epsilon$ *and* $\mathbf{A}$ *is separable [32]. Under these assumptions* Hottopix$(\tilde{\mathbf{X}}, m, \epsilon)$ *returns* $\hat{\mathbf{W}}$ *such that,*

$$\left\|\hat{\mathbf{W}} - \mathbf{W}_{:,S}\right\|_{\infty,1} \leq \epsilon \tag{24}$$

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

$$\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty,1} \leq \frac{4\epsilon}{\rho_1 - \epsilon} \tag{25}$$

Proof. Let $\mathbf{W}' = \mathbf{W}_{:,S}$ and $\mathbf{X} = \mathbf{A}\mathbf{W}_{:,S}$. The bound in (6) is immediate from Theorem 2 in [18] as $\mathbf{W}'$ is ρ_1-robust simplical by Theorem 6. We first note that,

$$\begin{aligned}\left\|\tilde{\mathbf{X}} - \mathbf{A}\hat{\mathbf{W}}\right\|_{\infty,1} &\leq \left\|\tilde{\mathbf{X}} - \mathbf{X}\right\|_{\infty,1} + \|X - \mathbf{A}\mathbf{W}'\|_{\infty,1} + \left\|\mathbf{A}\mathbf{W}' - \mathbf{A}\hat{\mathbf{W}}\right\|_{\infty,1} \\ &\leq \left\|\mathbf{A}\left(\mathbf{W}' - \hat{\mathbf{W}}\right)\right\|_{\infty,1} + \epsilon \\ &\leq \|\mathbf{A}\|_{\infty,1}\left\|\mathbf{W}' - \hat{\mathbf{W}}\right\|_{\infty,1} + \epsilon \leq 2\epsilon\end{aligned}$$

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

$$\left\|\mathbf{X} - \mathbf{A}\hat{\mathbf{W}}\right\|_{\infty,1} = \left\|\mathbf{A}\mathbf{W}' - \hat{\mathbf{A}}\hat{\mathbf{W}}\right\|_{\infty,1} \leq 3\epsilon. \tag{26}$$

We can apply triangle inequality to get,

$$\begin{aligned}\left\|\left(\mathbf{A} - \hat{\mathbf{A}}\right)\mathbf{W}'\right\|_{\infty,1} &\leq \left\|\mathbf{A}\mathbf{W}' - \hat{\mathbf{A}}\hat{\mathbf{W}}\right\|_{\infty,1} + \left\|\hat{\mathbf{A}}\left(\mathbf{W}' - \hat{\mathbf{W}}\right)\right\|_{\infty,1} \\ &\leq 3\epsilon + \left\|\hat{\mathbf{A}}\right\|_{\infty,1}\left\|\mathbf{W}' - \hat{\mathbf{W}}\right\|_{\infty,1} \\ &\leq 3\epsilon + \left(1 + \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty,1}\right)\epsilon \end{aligned} \tag{27}$$

In order to get the desired result we need to lower bound the L.H.S in (27). Note that rowsum $\left(\mathbf{A} - \hat{\mathbf{A}}\right) = \mathbf{0}$. Therefore we have,

$$\left\|\left(\mathbf{A} - \hat{\mathbf{A}}\right)\mathbf{W}'\right\|_{\infty,1} \geq \left\|\mathbf{A} - \hat{\mathbf{A}}\right\|_{\infty,1}\rho_1 \tag{28}$$

by definition. Combining (28) and (27) we get the required bound. □

A.9 Noisy Recovery of Extreme Points

In this section we assume that $\mathbf{A}$ satisfies the ℓ_2-WStRIP property with parameter $(\delta/L, \rho_2, m')$.

Lemma 7. *If $\mathbf{A}$ satisfies the ℓ_2-WStRIP property with parameter $(\delta/L, \rho_2, 2m')$ then the sets $\{S(1), \cdots, S(l+1)\}$ with $|S(i)| = 2m'$ satisfy,*

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

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 $\left\{\sigma_m(\mathbf{A}_{S(i),:}) < \rho_2\right\}$. Note that by virtue of ℓ_2-WStRIP each of these events is true with probability atmost δ/L. □

If the conditions of the above lemma are satisfied we will call the corresponding sets *good*. Recall the definition of $\hat{\mathbf{M}}_i(t)$. We will show that if $\hat{\mathbf{A}}(t)$ is close to $\mathbf{A}$ and the matrices $\hat{\mathbf{M}}_i(t)$ are sufficiently close to their means, then we recover $\mathbf{W}$ upto the same accuracy. Let us define $\mathbf{M}_i = \mathbb{E}\left[\hat{\mathbf{M}}_i(t)\right]$.

Lemma 8. *Suppose* $\mathbf{A}$ *satisfies the* ℓ_2*-WStRIP property and* $\{S(1), S(2), \cdots S(l+1)\}$ *are good in the sense of Lemma 7. Given that* $\left\|\hat{\mathbf{A}}(t) - \mathbf{A}\right\|_{\infty,1} \leq \epsilon_1$ *and* $\left\|\hat{\mathbf{M}}_i(t) - \mathbf{M}_i\right\|_{\infty,\infty} \leq \epsilon_2$ *for all* $i \in [l+1]$, $\hat{\mathbf{W}}(t)$ *recovered by Algorithm 1 satisfies,*

$$\left\|\hat{\mathbf{W}}(t) - \mathbf{W}\right\|_{\infty,\infty} \leq \frac{m(2\epsilon_1 + 3\epsilon_2)}{\rho_2} \tag{29}$$

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

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

$$\begin{aligned} \mathbf{A}_i\mathbf{W}_i &= \mathbf{M}_i \\ \hat{\mathbf{A}}_i(t)\hat{\mathbf{W}}_i(t) &= \hat{\mathbf{M}}_i(t) \end{aligned} \tag{30}$$

Note that $\mathbf{A}_i$ has full-column rank. Let the left-inverse of $\mathbf{A}_i$ be $\mathbf{A}_i^*$. It is easy to see that,

$$\|\mathbf{A}_i^*\|_{\infty,1} \leq \frac{m}{\rho_2}. \tag{31}$$

From (30) we have,

$$\begin{aligned} &\left(\mathbf{I} + \mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)\hat{\mathbf{W}}_i(t) = \mathbf{W}_i + \mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) \\ &\implies \hat{\mathbf{W}}_i(t) = \left(\mathbf{I} + \mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)^{-1}\left(\mathbf{W}_i + \mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i)\right) \\ &\implies \hat{\mathbf{W}}_i(t) = \left(\mathbf{I} - \mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)(\mathbf{I} + \mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i))\right)(\mathbf{W}_i + \mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i)) \end{aligned}$$

We can simplify further to yield,

$$\begin{aligned} \hat{\mathbf{W}}_i(t) - \mathbf{W}_i = \mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) - \Big(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\mathbf{W}_i + \left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)^2\mathbf{W}_i\Big) \\ - \Big(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) + \left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)^2\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i)\Big) \end{aligned}$$

Therefore by triangle inequality we have,

$$\begin{aligned} &\left\|\hat{\mathbf{W}}_i(t) - \mathbf{W}_i\right\|_{\infty,1} = \left\|\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i)\right\|_{\infty,1} \\ &+ \left\|\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\mathbf{W}_i + \left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)^2\mathbf{W}_i\right)\right\|_{\infty,1} \\ &+ \left\|\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i) + \left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t) - \mathbf{A}_i)\right)^2\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t) - \mathbf{M}_i)\right)\right\|_{\infty,1} \end{aligned}$$

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

$$\begin{aligned}\left\|\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t)-\mathbf{M}_i)\right\|_{\infty,1} &\leq \|\mathbf{A}_i^*\|_{\infty,1}\left\|(\hat{\mathbf{M}}_i(t)-\mathbf{M}_i)\right\|_{\infty}\\ &\leq \frac{m\epsilon_2}{\rho_2}\end{aligned}$$

Similarly we have,

$$\begin{aligned}&\left\|\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t)-\mathbf{A}_i)\mathbf{W}_i+\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t)-\mathbf{A}_i)\right)^2\mathbf{W}_i\right)\right\|_{\infty,1}\\ &\leq \|\mathbf{A}_i^*\|_{\infty,1}\left(1+\|\mathbf{A}_i^*\|_{\infty,1}\epsilon_1\right)\epsilon_1\|\mathbf{W}_i\|_{\infty,\infty}\\ &\leq \frac{2m\epsilon_1}{\rho_2}\end{aligned}$$

Finally the third term can be bounded as,

$$\begin{aligned}&\left\|\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t)-\mathbf{A}_i)\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t)-\mathbf{M}_i)\ +\left(\mathbf{A}_i^*(\hat{\mathbf{A}}_i(t)-\mathbf{A}_i)\right)^2\mathbf{A}_i^*(\hat{\mathbf{M}}_i(t)-\mathbf{M}_i)\right)\right\|_{\infty,1}\\ &\leq \left(\|\mathbf{A}_i^*\|_{\infty,1}\right)^2\epsilon_1\epsilon_2+\left(\|\mathbf{A}_i^*\|_{\infty,1}\right)^3\epsilon_1^2\epsilon_2 \leq \frac{2m\epsilon_2}{\rho_2}\end{aligned}$$

Therefore we have,

$$\left\|\hat{\mathbf{W}}_i(t)-\mathbf{W}_i\right\|_{\infty,1} \leq \frac{m(2\epsilon_1+3\epsilon_2)}{\rho_2}$$

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

A.10 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 ℓ_1-WStRIP and ℓ_2-WStRIP. 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 $\Delta = \min_{s\in[L]}\left(u^*(s) - \max_{k\neq k*(s)} U_{sk}\right)$ 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 $O\left(LK\log(t)\right)$ for general contextual bandit to only an $O\left(L\text{poly}(m,\log(K))\log(T)\right)$ dependence.

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

- $\beta = \min_s \beta_s = \Omega\left(1/L\right)$.

- $\mathbf{W} \in \mathbb{R}^{m \times K}$ *satisfies* ℓ_1*-WStRIP with parameters* $(\delta, \rho_1, 2m')$
- $\mathbf{A} \in [0,1]^{L \times m}$ *satisfies* ℓ_2*-WStRIP with parameters* $(\delta/L, \rho_2, 2m')$ *and is separable [32].*

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,*

$$\begin{aligned} R(T) &\leq \frac{\theta(m+2m')\log(T)}{\beta} + 4(L+K+1)m'\log(T) + o(1) \\ &= O\left(L\frac{\text{poly}(m,m')}{\Delta^2}\log T\right) \\ &= O\left(L\frac{m^5\log^2 K}{\Delta^2}\log T\right) \end{aligned}$$

where $\theta \geq 4\max\left(\frac{2m'((16+\Delta)\rho_2+32m)}{\Delta\rho_1\rho_2}, \frac{15}{\rho_1(1-\lambda)}\right)^2$.

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, $\left\|\hat{\mathbf{F}}(t) - \mathbf{F}\right\|_{\infty,\infty} \leq \epsilon_1(t)$ *and* $\left\|\hat{\mathbf{M}}_i(t) - \mathbf{M}_i\right\|_{\infty} \leq \epsilon_2(t)$ *for all* $i \in [l+1]$*. If the following conditions hold,*

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

then $k(t) = k^*(s_t)$*, 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,

$$\left\|\hat{\mathbf{A}}(t) - \mathbf{A}\right\|_{\infty,1} \leq \frac{8m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} \tag{33}$$

Since we have,

$$\begin{aligned} \epsilon_1(t) &\leq \frac{m\rho_1}{2m'(4\rho_2+m)} \\ \epsilon_2(t) &\leq \frac{\rho_2}{m} \end{aligned}$$

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

$$\left\|\hat{\mathbf{W}}(t) - \mathbf{W}(t)\right\|_{\infty,\infty} \leq \frac{m}{\rho_2}\left(\frac{16m'\epsilon_1(t)}{\rho_1 - 2m'\epsilon_1(t)} + 3\epsilon_2(t)\right) \tag{34}$$

Therefore we have,

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

Now, under the conditions of the lemma in (32), we have

$$\begin{aligned}
\frac{8m'\epsilon_1(t)}{\rho_1-2m'\epsilon_1(t)}\left(1+\frac{2m}{\rho_2}\right) &\leq \frac{\Delta}{4} \\
3\frac{m\epsilon_2(t)}{\rho_2} &\leq \frac{\Delta}{4}
\end{aligned}$$

This further implies that,

$$\left\|\hat{\mathbf{U}}(t)-\mathbf{U}\right\|_{\infty,\infty} \leq \frac{\Delta}{2}$$

This guarantees that we select the optimal arm at time-step t. □

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

Lemma 10. *Suppose* $\epsilon_t = \frac{(m+2m')\theta}{\beta t}$ *where* $\beta = \min_s \beta_s$. *Algorithm 1 ensures that,*

$$\begin{aligned}
\mathbb{P}\left(T_{sj}(t) < \frac{\theta}{2}H_t\right) &\leq \frac{1}{t^{\theta/12}} \\
\mathbb{P}\left(N^{(i)}(t)_{sj} < \frac{\theta}{2}H_t\right) &\leq \frac{1}{t^{\theta/12}}
\end{aligned}$$

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 $\mathbf{G}(0)$ to be sampled provide $E(t)=1$ and $H_t = 1$. Note that,

$$\begin{aligned}
\mathbb{E}\left[T_{sj}(t)\right] &\geq \sum_{l=1}^{t}\mathbb{P}\left(S_l = s, E(l)=1, H_l = 1, C_l = j\right) \\
&\geq \sum_{l=1}^{t}\frac{\theta}{l} = \theta H_t
\end{aligned}$$

Now, a straight forward application of Chernoff-Hoeffding's inequality yields,

$$\begin{aligned}\mathbb{P}\left(T_{sj}(t) < (1-\delta)\mathbb{E}\left[T_{sj}(t)\right]\right) &\leq \exp\left(-\frac{\delta^2}{3}\mathbb{E}\left[T_{sj}(t)\right]\right)\\ &\leq \exp\left(-\frac{\delta^2}{3}\theta H_t\right)\end{aligned}$$

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 $\mathbf{G}(i)$ to be sampled when $E(t) = 1$,$S_t = s_s(i)$ and $H_t = 0$.

$$\begin{aligned}\mathbb{E}\left[N^{(i)}(t)_{sj}\right] &\geq \sum_{l=1}^{t}\mathbb{P}\left(E(t)=1, S_t = s_s(i), H_t = 0, C'_l = j\right)\\ &\geq \sum_{l=1}^{t}\frac{\theta}{l} = \theta H_t\end{aligned}$$

□

The same concentration inequality as before applies.

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

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

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,

$$\begin{aligned}&\mathbb{P}\left(|\hat{\mathbf{F}}(t)_{sj} - \mathbf{F}_{sj}| > \epsilon_1(t)\right) \leq \mathbb{P}\left(|\hat{\mathbf{F}}(t)_{sj} - \mathbf{F}_{sj}| > \epsilon_1(t)\middle| T_{sj}(t) \geq \frac{\theta}{2}H_t\right) + \mathbb{P}\left(T_{sj}(t) < \frac{\theta}{2}H_t\right)\\ &\leq 2e^{-\frac{\epsilon_1(t)^2}{2}\frac{\theta\log(t)}{2}} + \frac{1}{t^{\theta/12}}\end{aligned}$$

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. □

Similarly we can bound the errors in estimating $\mathbf{M}_i$'s as in the lemma below.

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

$$\mathbb{P}\left(\cup_{i\in[l+1]}\left\{\left\|\hat{\mathbf{M}}_i(t) - \mathbf{M}_i\right\|_{\infty,\infty} > \epsilon_2(t)\right\}\right) \leq 4(K+1)m' \exp\left(-\frac{\epsilon_2(t)^2}{2}\frac{\theta\log(t)}{2}\right) + \frac{2(K+1)m'}{t^{\theta/12}}$$

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

$$\begin{aligned}&\mathbb{P}\left(|\hat{\mathbf{M}}_i(t)_{sj} - \mathbf{M_i}_{sj}| > \epsilon_1(t)\right) \leq \mathbb{P}\left(|\hat{\mathbf{M}}_i(t)_{sj} - \mathbf{M_i}_{sj}| > \epsilon_1(t)\middle| T_{sj}(t) \geq \frac{\theta}{2}H_t\right) + \mathbb{P}\left(T_{sj}(t) < \frac{\theta}{2}H_t\right)\\ &\leq 2e^{-\frac{\epsilon_2(t)^2}{2}\frac{\theta\log(t)}{2}} + \frac{1}{t^{\theta/12}}\end{aligned}$$

We can apply union bound over all the entries of all the $l+1$ matrices to get the result. □

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

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

$$\theta \geq 4 \max \left(\frac{2m'((16+\Delta)\rho_2 + 32m)}{\Delta \rho_1 \rho_2}, \frac{15}{\rho_1(1-\lambda)} \right)^2 \tag{35}$$

By virtue of the ℓ_1-WStRIP property of $\mathbf{W}$, the set S is ρ_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,

$$\begin{aligned} &\mathbb{P}\left(\left\| \hat{\mathbf{F}}(t) - \mathbf{F} \right\|_{\infty,\infty} > \frac{2}{\sqrt{\theta}} \right) \leq \frac{4Lm'}{t} + o\left(\frac{1}{t^2}\right) \\ &\mathbb{P}\left(\cup_{i \in [l+1]} \left\{ \left\| \hat{\mathbf{M}}_i(t) - \mathbf{M}_i \right\|_{\infty,\infty} > \frac{2}{\sqrt{\theta}} \right\} \right) \leq \frac{4(K+1)m'}{t} + o\left(\frac{1}{t^2}\right) \end{aligned} \tag{36}$$

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

$$R(T) \leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{E(t)=1\right\}\right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left\{E(t)=0\right\}\right] \mathbb{P}\left(k(t) \neq k^*(s_t)\right) \tag{37}$$

By Lemma 9 we have that,

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

We can combine this with (37) to get,

$$\begin{aligned} R(T) &\leq \frac{\theta(m+2m')\log(T)}{\beta} + 4(L+K+1)m'\log(T) + o(1) \\ &= O\left(L\text{poly}(m,m')\log(T)\right) \end{aligned}$$

if we assume that $1/\beta = O(L)$. □

A.11 Lower Bound for α-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 $\mathbf{U}$ and for α-consistent policies. For, any $z_i \in \mathcal{Z}$ we define $\mathcal{C}(z_i)$ as,

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

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

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

for any $T > \tau$, where C, τ are universal constants independent of problem parameters and $D(\mathbf{U})$ is a constant that depends on the entries of $\mathbf{U}$ 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 ([37])**.** *Consider two probability measures P and Q, both absolutely continuous with respect to a given measure. Then for any event $\mathcal{A}$ we have:*

$$P(\mathcal{A}) + Q(\mathcal{A}^c) \geq \frac{1}{2}\exp\{-\min(\mathrm{KL}(P||Q), \mathrm{KL}(Q||P))\}$$

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 $z_i \in \mathcal{Z}$ let us define $T(z_i) = \sum_{t=1}^{T} \mathbb{1}\{S_t \in \mathcal{C}(z_i)\}$. 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 $\mathcal{C}(z_i)$. By the assumptions in the theorem we have,

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

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

$$\mathbb{P}\left(\cup_{z_i \in \mathcal{Z}} E_i^c\right) \leq 2me^{-T/12} = o\left(\frac{1}{T^2}\right)$$

Fix a $z_i \in \mathcal{Z}$ and let k be the index of an arm that is not optimal for any of the contexts that belong to $\mathcal{C}(z_i)$. Let us create another system with parameter $(\mathbf{U}', \boldsymbol{A}, \mathbf{W}')$ where we make the entry $W_{ik} = \lambda = \frac{U_{max}+1}{2}$ 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 $s \in \mathcal{C}(z_i)$. Let A be the event defined as follows,

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

Now, in the system with parameter $\mathbf{U}$ for any $s \in \mathcal{C}(z_i)$ we have,

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

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

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

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

$$\begin{aligned}\mathbb{P}_{\mathbf{U}}(A) &\leq \frac{2C|\mathcal{C}(z_i)|^{1-\alpha}}{T(z_i)^{1-\alpha}} \\ \mathbb{P}_{\mathbf{U}'}(A^c) &\leq \frac{2(K-1)C|\mathcal{C}(z_i)|^{1-\alpha}}{T(z_i)^{1-\alpha}}\end{aligned} \tag{38}$$

Now from Lemma 13 we have,

$$\begin{aligned}&\mathrm{KL}\left(\mathbb{P}^T_{\mathbf{U}}, \mathbb{P}^T_{\mathbf{U}'}\right) \\ &\geq (1-\alpha)\left(\log(T(z_i)) - \log(L/m)\right) - \log(4KC)\end{aligned}$$

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

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

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,

$$\begin{aligned}R(T(z_i)) &\geq \Delta(z_i) \sum_{k\neq k^*(z_i)} \sum_{s\in\mathcal{C}(z_i)} \sum_{\{t:S_t=s\}} \mathbb{E}_{\mathbf{U}}\left[\{X_t = k\}\right] \\ &\geq \left(\operatorname*{argmin}_k \frac{(K-1)\Delta(z_i)}{\mathrm{KL}\left(U_{sk}, \lambda\right)}\right)\left((1-\alpha)\left(\log(T(z_i)) - \log(L/m)\right) - \log(4KC)\right)\end{aligned}$$

Let $D(\mathbf{U}) = \left(\operatorname{argmin}_{z_i,k} \frac{(K-1)\Delta(z_i)}{\mathrm{KL}(U_{sk},\lambda)}\right)$. Now, we have

$$\begin{aligned}R(T) &= \sum_{z\in\mathcal{Z}} \mathbb{E}\left[R(T(z_i))\right] \\ &\geq D(\mathbf{U})(K-1)\mathbb{E}\sum_{z\in\mathcal{Z}}\left((1-\alpha)\left(\log(T(z_i)) - \log(L/m)\right) - \log(4KC)\right)\end{aligned}$$

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

$$\begin{aligned}R(T) &= \sum_{z\in\mathcal{Z}} \mathbb{E}\left[R(T(z_i))\right] \\ &= \sum_{z\in\mathcal{Z}} \mathbb{E}\left[R(T(z_i))|E\right]\mathbb{P}(E) + \mathbb{E}\left[R(T(z_i))|E^c\right]\mathbb{P}(E^c) \\ &\geq D(\mathbf{U})(K-1)m\left((1-\alpha)\left(\log(T/2m) - \log(L/m)\right) - \log(4KC)\right) + o(1)\end{aligned}$$

□