JMLR: Workshop and Conference Proceedings vol 65:1–26, 2017 30th Annual Conference on Learning Theory

Semi-Bandits with Knapsacks

author names withheld

Editor: Under Review for COLT 2017

Abstract

This paper unifies two lines of work on multi-armed bandits: *Bandits with Knapsacks* (BwK) and *semi-bandits*. The former concerns scenarios when there are limited "resources" consumed by the algorithm, *e.g.,* limited inventory in a dynamic pricing problem. In the latter, there may be a huge number of actions, but there is combinatorial structure and additional feedback which makes the problem tractable. Both lines of work has received considerable recent attention, and are supported by numerous application examples.

We define a common generalization of the two, termed *Semi-Bandits with Knapsacks*. We design a general algorithm for this model, achieving regret rates that are comparable with those for BwK and semi-bandits in general, and in fact essentially optimal for each of the two special cases.

Keywords: Bandit algorithms, Semi-Bandit feedback, Randomized Rounding, Knapsacks

1. Introduction

Multi-armed bandits (*MAB*) is an elegant model for studying the tradeoff between acquisition and usage of information (a.k.a. *explore-exploit tradeoff*). In each round an algorithm sequentially chooses from a fixed set of alternatives (a.k.a. *actions*, a.k.a. *arms*), and receives reward for the chosen action. Crucially, the algorithm does not have enough information to answer all "counterfactual" questions ("what would have happened if a different action were chosen in this round").

MAB problems and various generalizations thereof is have been studied steadily since 1930-ies, with a huge surge of interest in the last decade. The work on MAB progresses across several directions, such as: what auxiliary information (if any) is revealed to the algorithm before and/or after it needs to make a decision, which "process" are the rewards are coming from, do they have some known structure that can be leveraged, are there global constraints on the algorithm, etc. Many of these directions gave rise to prominent lines of work.

This paper unifies two lines of work on MAB: *Bandits with Knapsacks* (BwK) and *semi-bandits*. The former concerns scenarios when there are limited "resources" consumed by the algorithm, *e.g.,* limited inventory in a dynamic pricing problem. In the latter, there may be a huge number of actions, but there is structure which makes the problem tractable. Namely, actions correspond to subsets of some "ground set", rewards are additive across the elements of this ground set (*atoms*), and the reward for each chosen atom is revealed after each round. This happens, *e.g.,* in an online routing problem, where each action is a path in a graph, *i.e.,* a (feasible) subset of edges. Both lines of work has received considerable recent attention, and are supported by numerous application examples.

Our contributions. We define a common generalization of the semi-bandits and BwK, termed *Semi-Bandits with Knapsacks* (SemiBwK). Following much of the literature, we focus on an i.i.d. environment: in each round, the "outcome matrix" — reward and resource consumption for all atoms — is

drawn independently from a fixed distribution over such matrices.[1] We design a general algorithm for this model. We achieve regret rates that are comparable with those for BwK and semi-bandits in general, and in fact essentially optimal for each of the two special cases.

Specifics of the result are as follows. As usual, rewards and resource consumption is bounded: in our setting, it means that for each atom, rewards and consumption of each resource in each round is at most 1. We define regret relative to OPT, the expected total reward of the best all-knowing policy. For BwK problems, OPT is known to be a much stronger benchmark than the traditional best-fixed-arm benchmark. We upper-bound regret in terms of the parameters of the problem: time horizon T, (smallest) budget B, number of atoms n, and OPT itself. We obtain

$$\texttt{Regret} \leq \tilde{O}(\sqrt{n/B}\,\mathrm{OPT} + \sqrt{n\,\mathrm{OPT}} + n). \tag{1.1}$$

This rate is optimal for the special case of BwK, up to polylog factors (under a very minor assumption that $\mathrm{OPT} \geq \Omega(\#\text{arms})$). In fact, this optimality holds in a very strong sense: for every given tuple of parameters (n, B, OPT) (Badanidiyuru et al., 2013). For the special case of semi-bandits, the paradigmatic scenario of interest is when each feasible action corresponds to a subset of size $k \leq n$. Then one can upper-bound $\mathrm{OPT} \leq kT$, and obtain regret $\tilde{O}(\sqrt{nk\,\mathrm{OPT}})$, which is known to be optimal for semi-bandits, up to polylog factors (Kveton et al., 2015b).

Our results hold under assumption that the action set, *i.e.,* the family of feasible subsets of atoms, is described by a *matroid constraint*. This is a rather general scenario which includes several paradigmatic special cases of semi-bandits such as cardinality constraints, source-sink path constraints, edge-disjoint paths constraints, spanning tree constraints, and matching constraints.

We list a number of applications and notable special cases of SemiBwK, and instantiate our results for these cases (Section 6).

Challenges and techniques. As observed in (Badanidiyuru et al., 2013), BwK problems are challenging compared to the traditional MAB problems with i.i.d. rewards because it no longer suffices to look for the best arm and/or optimize expected per-round rewards; instead, one essentially needs to look for a *distribution* over arms with optimal expected *total* reward across all rounds. Generic challenges in semi-bandits concern handling exponentially many actions (both in terms of regret and in terms of the running time), and taking advantage of the additional feedback. And in SemiBwK, one needs to deal not only with distributions over actions, but with distributions over subsets of atoms – which tend to be more complicated objects to deal with.

On a high level, we combine a technique from prior work on BwK with *randomized rounding* techniques from combinatorial optimization. We build on a BwK algorithm from (Agrawal and Devanur, 2014a), which combines linear relaxations and a well-known "optimism-under-uncertainty" paradigm. A generalization of this algorithm to SemiBwK results in a fractional solution $\boldsymbol{x}$ — a vector over the atoms — which needs to be converted to a feasible subset of atoms (or, more realistically, a distribution over feasible subsets). This is where randomized rounding techniques come in, constructing a distribution over feasible subsets that equals $\boldsymbol{x}$ in expectation. A notable conceptual challenge in our solution is to ensure that there is enough randomness in this distribution so that one can apply concentration bounds not only across rounds, but also across atoms. Technical challenges concern "opening up" the algorithm and analysis from (Agrawal and Devanur, 2014a) and connecting it with the insights from randomized rounding.

1. In particular, BwK problems have only studied under a similar i.i.d. assumption; BwK with adversarial rewards is terra incognita at this point.

As a tool, we rely on concentration bounds for negatively correlated random variables. In particular, we generalize a certain notion of "confidence radius" from independent random variables (for which it was previously used) to negatively correlated random variables.

Solving `SemiBwK` using prior work. Naively solving an instance of `SemiBwK` using an algorithm for `BwK` would result in a regret bound like (1.1) with n replaced with #actions, which could be on the order of n^k if each action can consist of at most k atoms, or perhaps even exponential in n.

`SemiBwK` can also be solved using (another) algorithm from (Agrawal and Devanur, 2014a) for a certain *linear* extension of `BwK` (termed "static contexts" in that paper). This algorithm exploits the combinatorial structure of actions, yet it ignores the additional feedback from the atoms. Agrawal and Devanur (2014a) only present analysis for the special case $B \geq \Omega(T)$, and for this special case achieve regret bound of the form $\tilde{O}(n\sqrt{T})$. Whereas we handle the full range of values for B, and for the directly comparable case $B \geq \Omega(T)$ our regret bound is better by a multiplicative factor $\sqrt{nT/\operatorname{OPT}}$, which can be large for problem instances when OPT is small.

Related work. Multi-armed bandits have been studied since Thompson (1933) in Operations Research, Economics, and several branches of Computer Science, see (Gittins et al., 2011; Bubeck and Cesa-Bianchi, 2012) for background. Among broad directions in MAB, most relevant is MAB with i.i.d. rewards, starting from (Lai and Robbins, 1985; Auer et al., 2002)

Bandits with Knapsacks (`BwK`) were first introduced by Badanidiyuru et al. (2013) as a common generalization of several models from prior work, as well as numerous motivating examples. Subsequent papers on `BwK` introduced new algorithms and "smoother" resource constraints Agrawal and Devanur (2014a), and generalized `BwK` to contextual bandits (Badanidiyuru et al., 2014; Agrawal et al., 2016; Agrawal and Devanur, 2016). Prior work on various special cases of `BwK` (*i.e.,* bandit problems with resources) includes dynamic pricing with limited supply (Babaioff et al., 2015; Besbes and Zeevi, 2009, 2012; Wang et al., 2014), dynamic procurement on a budget (Badanidiyuru et al., 2012; Singla and Krause, 2013; Slivkins and Vaughan, 2013), dynamic ad allocation with advertisers' budgets (Slivkins, 2013), and bandits with a single deterministic resource (Guha and Munagala, 2007; Gupta et al., 2011; Tran-Thanh et al., 2010, 2012).

The Semi-Bandit model was first studied by György et al. (2007) in the adversarial setting. In the i.i.d. setting, in a series of works by Gai et al. (2012), Chen et al. (2013) and finally by Kveton et al. (2015b) gave optimal algorithm for the Combinatorial Semi-Bandit problem. This was then extended to the linear Combinatorial Semi-Bandit setting by Wen et al. (2015). Kveton et al. (2014) studied the special case of Matroid Bandits. The other works in i.i.d. semi-bandits include Gabillon et al. (2013), Kveton et al. (2015a), Combes et al. (2015), Katariya et al. (2016), Zong et al. (2016) among others. Under semi-bandits with linear generalization, some works include Abbasi-Yadkori et al. (2011), Wen et al. (2015) among others. A common generalization of contextual bandits and semi-bandits has been studied in (Krishnamurthy et al., 2016).

Our algorithm uses the randomized rounding schemes from Combinatorial Optimization (Papadimitriou and Steiglitz (1982)) literature. These schemes were developed in the context of approximation algorithms (see Williamson and Shmoys (2011) for more on approximation algorithms). In particular, Raghavan and Tompson (1987) introduced the first randomized rounding scheme where they considered independent sampling with scaling. Gandhi et al. (2006), Asadpour et al. (2010), Chekuri et al. (2010), Chekuri et al. (2011) among others, developed randomized rounding schemes for combinatorial optimization problems, that used correlation among the rounded variable to give

sharp concentration bounds for linear combination of the rounded variables, for various approximation algorithms

Open questions. The most tempting open question is to consider a version in which each atom is characterized with a known low-dimensional feature vector, and its reward and resource consumption is a linear function of the features. Similar versions are common in the literature on semi-bandits (Wen et al., 2015), where to alleviate the dependence on #atoms by (fully or partially) replacing it with the dependence on #features.

If the feature vectors change from one round to another, this corresponds to contextual bandits, or rather a common generalization of contextual bandits, semi-bandits, and BwK. It is worth noting that prior work has combined *any two* of the three.

2. Our model and preliminaries

Our model, called *Semi-Bandits with Knapsacks* (SemiBwK) is a generalization of multi-armed bandits (henceforth, *MAB*) with i.i.d. rewards. As such, in each round $t = 1, \ldots, T$, an algorithm chooses an action S_t from a fixed set of actions $\mathcal{F}$, and receives a reward $\mu_t(S_t)$ for this action which is drawn independently from a fixed distribution that depends only on the chosen action. The number of rounds T, a.k.a. the *time horizon*, is known to the algorithm.

There are d resources being consumed by the algorithm. The algorithm starts out with budget $B_j \geq 0$ of each resource j. All budgets are known to the algorithm. If in round t action $S \in \mathcal{F}$ is chosen, the outcome of this round is not only the reward $\mu_t(S)$ but the consumption $C_t(S,j)$ of each resource $j \in [d]$. We refer to $\boldsymbol{C}_t(S) = (C_t(S,j) : j \in [d])$ as the *consumption vector*.

Actions correspond to subsets of a finite ground set $\mathcal{A}$, with $n = |\mathcal{A}|$; we refer to elements of $\mathcal{A}$ as *atoms*. Thus, the set $\mathcal{F}$ of actions corresponds to the family of "feasible subsets" of $\mathcal{A}$. The rewards and resource consumption is additive over the atoms: for each round t and each atom a there is a reward $\mu_t(a) \in [0,1]$ and consumption vector $\boldsymbol{C}_t(a) \in [0,1]^d$ such that for each action $S \subset \mathcal{F}$ it holds that $\mu_t(S) = \sum_{a \in S} \mu_t(a)$ and $\boldsymbol{C}_t(S) = \sum_{a \in S} \boldsymbol{C}_t(a)$.

We assume the i.i.d. property across rounds, but allow arbitrary correlations within the same round. Formally, for each atom a and each round t we define the *outcome vector* which includes both the reward and resource consumption:

$$v_t(a) := (\ \mu_t(a);\ C_t(a,1),\ \ldots,\ C_t(a,d)\ ) \in [0,1]^{d+1}.$$

We assume that the entire $n \times (d+1)$ matrix $(v_t(a) : a \in \mathcal{A})$ is chosen independently from a fixed distribution $\mathcal{D}_{\texttt{M}}$ over such matrices. The distribution $\mathcal{D}_{\texttt{M}}$ is not revealed to the algorithm.

Algorithm stops as soon as any one of the resources goes strictly below 0. The round in which this happens is called the stopping time and denoted $\tau_{\texttt{stop}}$. The reward in collected in this last round does not count; so the total reward of the algorithm is

$$\texttt{rew} = \textstyle\sum_{t < \tau_{\texttt{stop}}} \mu_t(S_t).$$

To recap, instance of SemiBwK consists of the action set $\mathcal{F} \subset 2^{[n]}$, the budgets $\boldsymbol{B} = (B_j : j \in [d])$, and the distribution $\mathcal{D}_{\texttt{M}}$. The $\mathcal{F}$ and and $\boldsymbol{B}$ are known to the algorithm, and $\mathcal{D}_{\texttt{M}}$ is not.

As explained in the introduction, SemiBwK subsumes *Bandits with Knapsacks* (BwK) and semi-bandits. Indeed, BwK is the special case when $\mathcal{F}$ consists of singletons, and semi-bandits is the special case when all budgets are $B_j = T$ (so that the resource consumption is irrelevant).

Regret. Following the prior work on BwK, we compete against the "optimal all-knowing algorithm": an algorithm that optimizes the expected total reward for a given problem instance; its expected total reward is denoted OPT. As observed in Badanidiyuru et al. (2013), OPT can be much larger (*e.g.,* factor of 2 larger) than the expected cumulative reward of the best action, for a variety of important special cases of BwK.

Our goal is to minimize *regret*, defined as OPT minus the total reward:

$$\texttt{Regret} := \text{OPT} - \texttt{rew}.$$

We strive to upper-bound regret in terms of the key parameters: the time horizon T, the optimal value OPT, the number of atoms n and the smallest budget $\min_{j\in[d]} B_j$. We may also factor in the smallest size of a feasible subset: $k := \min_{S\in\mathcal{F}} |S|$, particularly when $k \ll n$.

Additional assumptions and notation. We assume that the action set $\mathcal{F}$ satisfies the following property. Recall that $\mathcal{F}$ is given by a *combinatorial constraint*, *i.e.,* a family of subsets. Consider subsets of atoms as n-dimensional binary vectors; then $\mathcal{F}$ corresponds to a finite set of points in $\mathbb{R}^n$. We assume that the convex hull H of $\mathcal{F}$ forms a polytope in $\mathbb{R}^n$. In other words, there exists a set of linear constraints over $\mathbb{R}^n$ whose set of feasible *integral* solutions is $\mathcal{F}$. We call such combinatorial constraint $\mathcal{F}$ *linearizable*; the convex hull H is called the polytope *induced* by $\mathcal{F}$. The family of all linearizable combinatorial constraints is denoted with LIN. We consider several examples of linearizable combinatorial constraints, including cardinality constraints, matroid constraints, spanning trees, and matchings, see Appendix B for self-contained background.

Following prior work on BwK, we assume w.l.o.g. that all budgets are the same: $B_j = B$ for all resources $j \in [d]$, since we can divide all budgets by the smallest budget $\min_{j\in[d]} B_j$.

The mean rewards and mean consumption is denoted $\mu(a) := \mathbb{E}[\mu_t(a)]$ and $\boldsymbol{C}(a) := \mathbb{E}[\boldsymbol{C}_t(a)]$. We extend the notation to subsets of atoms: $\mu(S) := \sum_{a\in S} \mu(a)$ and $\boldsymbol{C}(S) := \sum_{a\in S} \boldsymbol{C}(a)$.

2.1. Concentration Bounds

We use several concentration bounds, focusing on negatively correlated random variables.

Throughout this subsection, let $\mathcal{X} = (X_1, X_2, \ldots, X_m)$ denote a collection of random variables which take values in $[0, 1]$. Let $X := \frac{1}{m}\sum_{i=1}^{m} X_i$ be the average, and $\mu := \mathbb{E}[X]$ be its expectation.

Family $\mathcal{X}$ is called *negatively correlated* iff

$$\mathbb{E}\left[\prod_{i\in S} X_i\right] \leq \prod_{i\in S} \mathbb{E}[X_i] \quad \forall S \subseteq [m], \tag{2.1}$$

$$\mathbb{E}\left[\prod_{i\in S}(1 - X_i)\right] \leq \prod_{i\in S} \mathbb{E}[1 - X_i] \quad \forall S \subseteq [m], \tag{2.2}$$

Note that independent random variables are negatively correlated.

Claim 2.1 *If family $\mathcal{X}$ is negatively correlated, then $\left(\frac{1+|X_i-\mu_i|}{2} : i \in [m]\right)$ is, too.*

While this claim is probably known, we provide the proof in Appendix A for completeness.

We use many versions of the Chernoff-Hoeffding bounds for negatively correlated random variables. These can be found in Theorem 3.2 and Corollary 3.3 of Panconesi and Srinivasan (1992), Theorem 3.3 of Impagliazzo and Kabanets (2010).

Theorem 2.2 (Panconesi and Srinivasan (1992)) *If family $\mathcal{X}$ is negatively correlated then:*

(a) $\Pr[|X-\mu| > \epsilon\mu] \leq c\ \exp(-\mu m\epsilon^2/3)$ *for any* $0 < \epsilon < 1$.
(b) $\Pr[X > a] \leq 2^{-am}$ *for any* $a > 6\mu$.

Theorem 2.3 (Impagliazzo and Kabanets (2010)) *Suppose $\mathcal{X}$ is a family of negatively correlated random variables such that for every $i \in [m]$, $\mathbb{E}[X_i] = 1/2$. Then we have, $\Pr[X \geq 1/2 + \epsilon] \leq c.e^{-m\mathcal{D}_{\mathrm{KL}}(1/2+\epsilon \,\|\, 1/2)}$, where $\mathcal{D}_{\mathrm{KL}}(. \,\|\, .)$ is the KL divergence function.*

Following prior work (Kleinberg et al., 2015; Babaioff et al., 2015; Badanidiyuru et al., 2013; Agrawal and Devanur, 2014b), we use the following notion of *confidence radius*:

$$\mathrm{Rad}_\alpha(x, m) = \sqrt{\alpha x/m} + \alpha/m. \tag{2.3}$$

This notion is useful because if random variables $\mathcal{X}$ are independent, then

$$\Pr\left[\ |X-\mu| < \mathrm{Rad}_\alpha(X, m) < 3\,\mathrm{Rad}_\alpha(\mu, m)\right] > 1 - O\left(2^{-\Omega(\alpha)}\right), \quad \forall \alpha > 0. \tag{2.4}$$

Thus, we can bound the deviations $|X-\mu|$ in a way that gets sharper when the μ is small, without knowing the μ in advance. We generalize this result. Before stating our generalization, let us make some definition.

Let $\mathcal{Z} = \{Z_{t,a} : t \in [T], a \in [n]\}$ denote a family of random variables in $[0,1]$. Let $\mathcal{Z}_t := \{Z_{t',a} : a \in [n],\ \forall t' \leq t\}$. Suppose, $\{M_{t,a} : t \in [T], a \in [n]\}$ denotes a set of multipliers in $[0,1]$ such that $M_{t,a}$ is completely determined given $\mathcal{Z}_{t-1}$. Let $\hat{\mu} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{a=1}^{n}\mathbb{E}[M_{t,a}Z_{t,a} \mid \mathcal{Z}_t - 1]$ and let $Z = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} M_{t,a}Z_{t,a}$. Suppose we have Theorem 2.2 part (a) in the following form.

$$\begin{aligned} \text{for any } \epsilon > 0 \quad & \Pr[(Z-\hat{\mu}) > \epsilon\hat{\mu}] \leq \left(\frac{e^\epsilon}{(1+\epsilon)^{1+\epsilon}}\right)^{nT\hat{\mu}} \leq c\ \exp(-\hat{\mu}nT\epsilon^2/3) \\ \text{for any } 0 < \epsilon < 1 \quad & \Pr[(Z-\hat{\mu}) > \epsilon\hat{\mu}] \leq c\ \exp(-\hat{\mu}nT\epsilon^2/3) \end{aligned} \tag{2.5}$$

It is well-known (we provide a proof in Appendix for completeness) that (2.5) can be transformed into the following form

$$\Pr[Z > a] \leq 2^{-anT} \text{ for any } a > 6\hat{\mu} \tag{2.6}$$

Theorem 2.4 *If family $\mathcal{X}$ is negatively correlated, then* (2.4) *holds. Additionally, for a family $\mathcal{Z}$ as defined above, if* (2.5) *holds, then* (2.4) *holds for Z with μ replaced by $\hat{\mu}$ and m replace by nT.*

The proof builds on Theorem 2.2(ab) and is very similar to the one for independent random variables; we provide it in Appendix A for completeness.

2.2. Confidence bounds

We use the confidence radius (2.3) to define standard upper/lower confidence bounds on the mean rewards and mean consumption. Fix round t, atom a, and resource j. Let $\hat{\mu}_t(a)$ and $\hat{C}_t(a,j)$ denotes the empirical average of the rewards and resource-j consumption, resp., between rounds 1 and $t-1$. Let $N_t(a)$ be the number of times atom a has been chosen in these rounds (*i.e.,* how many times it has been included in the chosen action). Fixing parameter $\alpha > 0$ to be chosen later, the upper/lower confidence bounds are defined as

$$\begin{aligned} \mu_t^{\pm}(a) &= \texttt{proj}\,(\quad \hat{\mu}(a) \pm \mathrm{Rad}_\alpha(\hat{\mu}(a), N_t(a)) \quad) \\ C_t^{\pm}(a,j) &= \texttt{proj}(\quad \hat{C}(a,j) \pm \mathrm{Rad}_\alpha(\hat{C}(a,j), N_t(a)) \quad), \end{aligned} \tag{2.7}$$

where $\texttt{proj}(x) := \mathrm{argmin}_{y\in[0,1]} |y-x|$ denotes the projection into $[0,1]$. We use a vector notation $\boldsymbol{\mu}_t^{\pm}$ and $\boldsymbol{C}_t^{\pm}(j)$ to denote the corresponding n-dimensional vectors over all atoms a.

As an easy corollary of Theorem 2.4 (see Corollary A.1 for a more specific formulation), it follows that with probability $1 - O(e^{-\Omega(\alpha)})$ we have the following:

$$\begin{aligned} \mu_t^+(a) \geq \mu(a) \geq \mu_t^-(a) \\ C_t^+(a,j) \geq C(a,j) \geq C_t^-(a,j) \end{aligned}$$

In what follows, we always use the confidence radius with the same parameter α, to be specified later, which we suppress from the notation.

2.3. Randomized Rounding

We incorporate prior work on randomized rounding for linear programs. As applied to our setting, randomized rounding means the following. Assume action set $\mathcal{F}$ is linearizable, and consider the induced polytope $P \subset [0,1]^n$. The *randomized rounding scheme* (henceforth, RRS) for $\mathcal{F}$ is an algorithm that inputs a feasible fractional solution $\boldsymbol{x} \in P$ and the linear equations describing P, and produces a random vector $\boldsymbol{Y}$ over $\mathcal{F}$.

For our main result, we consider RRS's such that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{x}$ and $\boldsymbol{Y}$ is negatively correlated; we call such RRS's *negatively correlated*. Several examples of such RRS have been designed in literature: for cardinality constraints and bipartite matching (Gandhi et al., 2006), for spanning trees (Asadpour et al., 2010), and for matroids (Chekuri et al., 2010).

3. Main algorithm: SemiBwK-UCB

Let us define our main algorithm, called SemiBwK-UCB. The algorithm builds on an arbitrary RRS for the action set $\mathcal{F}$. It is parameterized by this RRS, the polytope $\mathcal{P}$ induced by $\mathcal{F}$ (represented as a collection of linear constraints), and a number $\epsilon > 0$. In each round t, it recomputes the upper/lower confidence bounds, as defined in Section 2.2, and solves the following linear program:

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{\mu_t^+} \cdot \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{C}_t^-(j) \cdot \boldsymbol{x} \leq \tfrac{B(1-\epsilon)}{T}, \quad j = 1, ..., d \\ & \boldsymbol{x} \in \mathcal{P} \\ & \boldsymbol{x} \in [0,1]^n. \end{aligned} \tag{$\mathrm{LP}_{\mathrm{ALG}}$}$$

This linear program defines a linear relaxation of the original problem which is "optimistic" in the sense that it uses upper confidence bounds for rewards and lower confidence bounds for consumption. The linear relaxation is also "conservative" in the sense that it rescales the budget by $1-\epsilon$. Parameter ϵ will be fixed throughout. For ease of notation, we will denote $B_\epsilon := (1-\epsilon)B$ henceforth. The LP solution $\boldsymbol{x}$ can be seen as a probability vector over the atoms.

Finally, the algorithm uses the RRS to convert the LP solution into a feasible action. The pseudocode is given as Algorithm 1.

Algorithm 1: SemiBwK-UCB

input: an RRS for action set $\mathcal{F}$, induced polytope $\mathcal{P}$ (as a set of linear constraints), $\epsilon > 0$.

for $t = 1, 2, \ldots, T$ **do**

1. **Recompute Confidence Bounds** according to (2.7)
2. **Obtain fractional solution** $\boldsymbol{x}_t \in [0,1]^n$ by solving LP_{ALG}.
3. **Obtain a feasible action** $S_t \in \mathcal{F}$ by invoking the RRS on $\boldsymbol{x}_t$.
4. **Semi-bandit Feedback**: observe the rewards/consumption for all atoms $a \in S_t$.

end

If action set $\mathcal{F}$ is described by a matroid constraint (see Appendix B for background), and we can use the RRS from Chekuri et al. (2010). In particular, we obtain a complete algorithm for several combinatorial constraints commonly used in the literature on semi-bandits, such as spanning tree constraints and shortest paths constraints.

Running time (for matroid constraints). At each round t, the algorithm does two computationally intensive steps: solves the linear program and runs the RRS.

The LP is on n variables and $O(2^n)$ constraints. Matroids are known to admit a polynomial-time seperation oracle (*e.g.,* see Schrijver, 2002)). It follows that the entire set of constraints in LP_{ALG} admits a polynomial-time separation oracle, and therefore we can then use Ellipsoid algorithm to solve LP_{ALG} in polynomial time. Moreover, for some specialized classes of matroids the LP is much smaller: *e.g.,* for cardinality constraints (just $d+1$ constraints) and for bipartite matching contraints (just $2n+d$ constraints). Then faster (near-linear-time) algorithms can be used.

The RRS from Chekuri et al. (2010) has $O(n^2)$ running time.[2] Hence, the computational bottleneck is usually solving the LP, rather than the RRS.

4. Regret Analysis: negatively correlated RRS

Let us analyze regret if the RRS is negatively correlated. Our main technical result is as follows:

Theorem 4.1 *Consider the* `SemiBwK` *problem with a linearizable action set $\mathcal{F}$ that admits a negatively correlated* RRS. *Then algorithm* SemiBwK-UCB *with this* RRS *and an appropriately*

2. We note in passing that for specialized classes of matroids, there are other RRS which run in linear time: *e.g.,* for the cardinality constraint (Gandhi et al., 2006), the spanning tree constraints (Asadpour et al., 2010), etc.

chosen parameter ϵ achieves

$$\texttt{Regret} \leq O(\log(ndT/\delta)) \ \left(\mathrm{OPT}\sqrt{n/B} + \sqrt{n\ \mathrm{OPT}} + n\right) \tag{4.1}$$

with probability at least $1-\delta$, for any $\delta > 0$. Here T is the time horizon, n is the number of atoms, and B is the budget. The parameter ϵ should be set to $\epsilon = \sqrt{\frac{\alpha n}{B}} + \frac{\alpha n}{B}$, where $\gamma = \log(ndT/\delta)$.

Corollary 4.2 *Suppose the action set $\mathcal{F}$ is defined by a matroid on the set of atoms. Then, using the negatively correlated* RRS *from (Chekuri et al., 2010), we obtain the regret in equation (4.1).*

Below we overview the proof of Theorem 4.1. Missing details can be found in Section 7.

4.1. Linear programs

We argue that $\mathrm{LP}_{\mathrm{ALG}}$ provides a good benchmark that we can use instead of OPT. Specifically, fix round t and let $\mathrm{OPT}_{\texttt{ALG},\,t}$ denote the optimal value for $\mathrm{LP}_{\mathrm{ALG}}$ in this round. Then:

Lemma 4.3 $\mathrm{OPT}_{\texttt{ALG},\,t} \geq \frac{1}{T}(1-\epsilon)\,\mathrm{OPT}$.

We will prove this by constructing a series of LP's, starting with a generic linear relaxation for BwK and ending with $\mathrm{LP}_{\mathrm{ALG}}$. We show that along the series the optimal value does not decrease.

The first LP, adapted from Badanidiyuru et al. (2013), has one decision variable for each action, and applies generically to any BwK problem.

$$\begin{array}{ll} \text{maximize} & \sum_{S\in\mathcal{F}}\ \mu(S)\,x(S) \\ \text{subject to} & \sum_{S\in\mathcal{F}}\ C(S,j)\,x(S) \leq B/T \quad j=1,...,d \\ & 0 \leq \sum_{S\in\mathcal{F}}\ x(S) \leq 1 \end{array} \tag{$\mathrm{LP}_{\texttt{BwK}}$}$$

Let $\mathrm{OPT}_{\texttt{BwK}}(B)$ denote the optimal value of this LP with a given budget B. Then:

Claim 4.4 $\mathrm{OPT}_{\texttt{BwK}}(B_\epsilon) \geq (1-\epsilon)\,\mathrm{OPT}_{\texttt{BwK}}(B) \geq \frac{1}{T}(1-\epsilon)\ \mathrm{OPT}$.

The second inequality in Claim 4.4 follows from (Lemma 3.1 in Badanidiyuru et al., 2013).

Now consider a simpler LP where the decision variables correspond to atoms. As before, $\mathcal{P}$ denotes the polytope induced by action set $\mathcal{F}$.

$$\begin{array}{ll} \text{maximize} & \boldsymbol{\mu}\cdot\boldsymbol{x} \\ \text{subject to} & C^\dagger\cdot\boldsymbol{x} \preccurlyeq B_\epsilon/T \quad \boldsymbol{x}\in\mathcal{P} \quad \boldsymbol{x}\in[0,1]^n. \end{array} \tag{$\mathrm{LP}_{\mathrm{ATOMS}}$}$$

Here $C = (C(a,j):\ a\in A, j\in d)$ is the $n\times d$ matrix of expected consumption, and $C^\dagger$ denotes its transpose. The notation $\preccurlyeq$ means that the inequality $\leq$ holds for for each coordinate.

Leting $\mathrm{OPT}_{\texttt{atoms}}$ denote the optimal value for $\mathrm{LP}_{\mathrm{ATOMS}}$, we have:

Claim 4.5 $\mathrm{OPT}_{\texttt{ALG},\,t} \geq \mathrm{OPT}_{\texttt{atoms}} \geq \mathrm{OPT}_{\texttt{BwK}}(B_\epsilon)$

Hence, combining Claim 4.4 and Claim 4.5, we obtain Lemma 4.3.

4.2. "Clean events"

Let us set up several events (henceforth called "clean events") and prove that they hold with high probability. Then the remainder of the analysis can proceed conditioning on these events. At a high-level, we use the same clean events as those used by Agrawal and Devanur (2014b) but use a somewhat different analysis; in particular, we will use properties of the RRS.

As a tool, we will need the following generalization of the concentration bound for negatively correlated random variables (Theorem 2.2):

Theorem 4.6 *Let $\mathcal{Z}_T = \{\zeta_t(a) : a \in [n], t \in [T]\}$ be a family of random variables taking values in $[0,1]$. Assume random variables $\{\zeta_t(a) : a \in [n]\}$ are negatively correlated given $\mathcal{Z}_{t-1}$ and have expectation $\frac{1}{2}$ given $\mathcal{Z}_{t-1}$, for each round t. Then family $\mathcal{Z}_T$ is negatively correlated.*

In what follows, it is convenient to consider a version of `SemiBwK` in which the algorithm does not stop, so that we can argue about what happens w.h.p. if our algorithm runs for the full T rounds. Then we show that our algorithm does indeed run for the full T rounds w.h.p.

Recall that $\boldsymbol{x_t}$ be the optimal fractional solution obtained by solving the LP in round t. Let $\boldsymbol{Y_t} \in \{0,1\}^n$ be the random binary vector obtained by invoking the RRS (so that the chosen action $S_t \in \mathcal{F}$ corresponds to a particular realization of $\boldsymbol{Y_t}$, interpreted as a subset).

"Clean event" for rewards. For brevity, for each round t let $\boldsymbol{\mu}_t = (\mu_t(a) : a \in A)$ be the vector of realized rewards, and let $r_t := \mu_t(S_t) = \boldsymbol{\mu}_t \cdot \boldsymbol{Y_t}$ be the algorithm's reward at this round.

Lemma 4.7 *Consider `SemiBwK` without stopping. Then with probability at least $1 - nT\ O(e^{-\Omega(\alpha)})$:*

$$\left|\sum_{t\in[T]} r_t - \sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}\right| \leq O\left(\sqrt{\alpha n \sum_{t\in[T]} r_t} + \alpha n\right).$$

Proof Sketch We will split the proof of this claim, by proving the following three equations.

With probability at least $1 - nT\ O(e^{-\Omega(\alpha)})$: the following holds:

$$\left|\sum_{t\in[T]} r_t - \sum_{t\in[T]} \boldsymbol{\mu} \cdot \boldsymbol{Y_t}\right| \leq 3nT \operatorname{Rad}\left(\tfrac{1}{nT}\sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}\ ,\ nT\right) \tag{4.2}$$

$$\left|\sum_{t\in[T]} \boldsymbol{\mu} \cdot \boldsymbol{Y_t} - \boldsymbol{\mu_t^+} \cdot \boldsymbol{Y_t}\right| \leq 12\sqrt{\alpha n \left(\sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}\right)} + 12\sqrt{\alpha} n + 12\alpha n \tag{4.3}$$

$$\left|\sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{Y_t} - \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}\right| \leq 3nT \operatorname{Rad}\left(\tfrac{1}{nT}\sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}\ ,\ nT\right) \tag{4.4}$$

We will use the properties of RRS to prove Equation (4.4). Proof of Equation (4.3) is similar to Agrawal and Devanur (2014b), while proof of Equation (4.2) follows immediately from the setup of the model. Using the parts (4.2) and (4.4) we can now find an appropriate upper bound on $\sqrt{\sum_{t\in[T]} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}}$ and using this upper bound, we prove Lemma 4.7.

We prove Equation (4.4) as follows. Recall that $\mu_t^+(a)$ is determined by the random variables $\mathcal{F}_{t-1} := \{Y_{t'}(a') : \forall t' < t, \forall a \in [n]\}$. Additionally note that, conditioned on a realization of $\mathcal{F}_{t-1}$, we have that the random variables $\{Y_t(a) : a \in [n]\}$ are negatively correlated. Hence, the random variables $\{\mu_t^+(a) Y_t(a) : a \in [n]\}$ are negatively correlated in this conditional space. Define $\tilde{\zeta}_t(a) = \mu_t^+(a) Y_t(a)$ for all $t \in [T], a \in [n]$. Note that $\mathbb{E}[\tilde{\zeta}_t(a) | \mathcal{F}_{t-1}] = \mu_t^+(a) x_t(a)$. Define $\zeta_t(a) = (1 + (\tilde{\zeta}_t(a) - \mu_t^+(a) x_t(a)))/2$. From Claim 2.1, we have that $\{\zeta_t(a) : a \in [n]\}$ conditioned on $\mathcal{F}_{t-1}$ are negatively correlated.

Note that, the family $\{\zeta_t(a) : t \in [T], a \in [n]\}$ satisfy the assumptions in Theorem 4.6. Hence we have that these random variables are negatively correlated. As a consequence, they satisfy Theorem 2.3. Hence, we get the standard additive form of the Chernoff-bounds for the random variable $\tilde{\zeta}_t(a)$ (*i.e.,* $\Pr[\frac{1}{nT}(\sum_{t=1}^{T}\sum_{a=1}^{n}\tilde{\zeta}_t(a) - \mu_t^+(a)x_t(a)) \geq \epsilon] \leq c.e^{-nT\mathcal{D}_{\mathrm{KL}}(1/2+\epsilon\,\|\,1/2)}$). Similarly, to get the lower tail of the additive form, consider a symmetric argument by letting $\zeta'_t(a) = (1 + (\mu_t^+(a)x_t(a) - \tilde{\zeta}_t(a)))/2$. Finally, it is well-known that this additive form can be transformed into the multiplicative form to obtain Eq. (2.5) (*e.g.,* see Mitzenmacher and Upfal (2005), Dubhashi and Panconesi (2009)). Hence, this implies that from Theorem 2.4 for the random variables $\{\tilde{\zeta}_t(a) : t \in [T], a \in [n]\}$ we have Eq. (2.4). Hence, we have that with probability $1 - \Omega(\exp(-\alpha))$ we have (4.4). ■

"Clean event" for consumption. We define a similar "clean event" for consumption of each resource j. The proof is similar and is deferred to later in this paper.

By a slight abuse of notation, for each round t let $\boldsymbol{C_t(j)} = (C_t(a,j) : \ a \in \mathcal{A})$ be the vector of realized consumption of resource j. Let $\chi_t(j)$ denote algorithm's consumption for resource j at time-step t (*i.e.,* $\chi_t(j) = \boldsymbol{C_t(j)} \cdot \boldsymbol{Y_t}$).

Lemma 4.8 *Consider* SemiBwK *without stopping. Then with probability at least* $1-nT\ O(e^{-\Omega(\alpha)})$*:*

$$\forall j \in [d] \quad |\sum_{t\in[T]} \chi_t(j) - \sum_{t\in[T]} \boldsymbol{C_t^-(j)} \cdot \boldsymbol{x_t}| \leq \sqrt{\alpha n B_\epsilon} + \alpha n.$$

4.3. Putting it all together

Similar to Agrawal and Devanur (2014b), we will handle the hard constraint on budget, by choosing an appropriate value of ϵ. We then combine the above Lemma on "rewards" clean event to compare the reward of the algorithm with that of the optimal value of LP to obtain the regret bound in Equation 4.1. Additionally, we use the Lemma on "consumption" clean event to argue that the algorithm doesn't exhaust the resource budget before time-step T.

5. Extension: arbitrary linearizable action set

We can extend SemiBwK-UCB to an arbitrary linearizable action set $\mathcal{F}$, assuming each resource is consumed by at most one atom (*e.g.,* this is the case for "Dynamic Assortment" problem, see Section 6). We use a very simple RRS: given a fractional solution $\boldsymbol{x}$ which lies in $\mathcal{P}$, the polytope induced by $\mathcal{F}$, we represent $\boldsymbol{x}$ as a distribution $\boldsymbol{Y}$ over the vertices of $\mathcal{P}$, and output $\boldsymbol{Y}$. This is a valid RRS because vertices of $\mathcal{P}$ lie in $\mathcal{F}$. However, while we get $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{x}$, we cannot guarantee negative correlation or any other similarly useful property.

Using analysis similar to Theorem 4.1,[3] we obtain:

Theorem 5.1 *Consider the* SemiBwK *problem with a linearizable action set. Assume each resource can be consumed by at most one atom. Use the same notation and same parameter* ϵ *as in Theorem 4.1. Then algorithm* SemiBwK-UCB *with the simple* RRS *described above achieves:*

$$\texttt{Regret} \leq O(\log(ndT/\delta)) \left(\mathrm{OPT}\sqrt{n/B} + n\sqrt{\mathrm{OPT}} + n\right). \tag{5.1}$$

with probability at least $1-\delta$*, for any* $\delta > 0$*.*

3. The main modification is a simpler-to-prove but less efficient version of Lemma 4.8.

The regret bound improves over the regret bound from (Agrawal and Devanur, 2014a) (discussed in the intro), when $B = o(T)$ or $\mathrm{OPT} = o(T)$, and coincides with that regret bound otherwise.

6. Applications and special cases

We discuss several notable applications and special cases of `SemiBwK`. Essentially, we generalize some of the numerous applications listed Badanidiyuru et al. (2013).

Dynamic Pricing with Multiple Items. The problem is as follows. The algorithm has d products on sale with limited supply of each: B_i units of product i. For simplicity, assume $B_i = B$ for all i. In each round t, an agent arrives. The algorithm chooses a vector of prices $(p_t(1), p_t(2), \ldots, p_t(d)) \in [0,1]^d$ to offer the agent. For simplicity, say, the agent is interested in buying (and/or is only allowed to buy) at most one item of each product. The agent has a valuation function $v : 2^{[d]} \to [0, d]$ such that they choose to buy the subset of items I that maximizes $v(I) - \sum_{i \in I} p_t(i)$. The valuation function can be arbitrary and not necessarily additive. As in prior work, we will assume that it is drawn as an independent sample at time t, from a fixed and unknown distribution.

We will frame this problem in `SemiBwK` framework. To side-step discretization issues, assume the algorithm can only choose prices form a given finite set $S \subset [0,1]$. The atoms then correspond to the (price, product) pairs. The constraint is that exactly one price is chosen for each product (which forms a partition matroid, see Appendix B). Resources correspond to products in an obvious way. Note that, in the model of `SemiBwK`, the entries rewards and consumptions matrix across atoms can be arbitrarily correlated, which helps us handle arbitrary agent valuations.

We consider S-regret: relative to OPT that is also restricted to prices in S. Note that $\mathrm{OPT} \leq dB$, since that is the maximum number of products available. Hence, the S-regret is $\tilde{O}(d\sqrt{dB|S|})$. This is an almost *exponential* improvement in the number of products compared to `BwK` framework. Recall that in that framework, arms correspond to every possible realization of prices for the d products. Hence, using our notation, the regret obtained is $\tilde{O}(d\sqrt{B|S|^d})$.

We will now obtain a regret bound for this problem using the linear `BwK` framework of Agrawal and Devanur (2014a). Note that their bound is instance independent. Hence, the regret they obtain is $\tilde{O}(d^2|S|\sqrt{T})$. When $B = O(T)$, our bound improves over this bound in two ways. Firstly, we have a $\sqrt{|S|}$ dependence as opposed to S. Note that S is a finite subset of possible prices and is usually large. Secondly, our bound also improves the dependence on the number of products. Additionally, when $B = o(T)$, their bounds do not apply.

Dynamic Assortment. This problem is similar to Dynamic Pricing in that the algorithm is selling products to an agent. However, the prices for these items are externally fixed. Instead, the algorithm has a large number of products and can offer only a limited number of those per time-step. Formally, the algorithm has d products. It has B_i copies of product i and for simplicity we will assume that all B_i are same and equal to B. At each time-step an agent arrives and the algorithm has to choose a subset of at most k products to show to the agent. Like in the case of Dynamic pricing, the agent has a valuation function over subsets and this valuation function is sampled each time from a fixed but unknown distribution. The agent will buy a subset of items based on their valuation function (i.e. the subset I among the k products shown that maximizes $v(I) - \sum_{i \in I} p_t(i)$). The goal of the algorithm is to sell as many products as possible in T time-steps.

To frame this as a `SemiBwK` problem: there is an atom for each of the d products. The resources correspond to the d products. The combinatorial constraint on the actions is the cardinality con-

straint. When an agent arrives, the algorithm can choose any subset of k atoms. The reward for an atom i is 1 if the agent buys product i and 0 otherwise. Similarly, the consumtion for resource i is 1 if the agent buys product i and 0 otherwise. We will now derive the regret using the SemiBwK-UCB algorithm. Note that $\mathrm{OPT} \leq dB$ and number of atoms is d. Hence, regret is $\tilde{O}(d\sqrt{dB})$.

Applying the BwK framework, we would have arms corresponding to each subset of k products. Hence, the number of arms would be $O(d^k)$. The other parameters of the problem would remain the same. Hence, the regret obtained would be $\tilde{O}(d\sqrt{Bd^k})$. Hence, again SemiBwK-UCB almost exponentially improves the regret bound. Note that, for this problem one can consider other types of constraints. For example, it is common that certain products naturally are sold together, or certain other products are naturally sold in a mutually exclusive manner. We can apply SemiBwK-UCB with Natural RRS under a large class of such constraints and obtain a similar bound.

Adjustable Prices in repeated multiple auctions. We consider the natural generalization of the repeated auctions with adjustable parameters (*e.g.,* , repeated second-price auction with adjustable reserve price (Cesa-Bianchi et al., 2013)). In this setting, the auctioneer is running multiple simultaneous repeated auctions (let r denote the number of auctions) to sell d items and trying to learn an estimate of some adjustable parameter (such as reserve prices in second-price auctions) through the various rounds of the auctions. The items are shared across all the auctions. We will assume that there are B_j copies of item j and as before, simplify it to $B_j = B$ for all items. Similar to previous work, we assume that in every round a fresh set of participants arrive with number of participants and the vector of their types in each auction being unknown but sampled independently from a fixed joint distribution at each time-step. We will assume that this adjustable parameter comes from a finite domain $S \subset [0, 1]$. Note that, at each time-step the algorithm fixes a value to this parameter in each of the r auctions. Then a random sample of participants types are drawn from a joint distribution across the auctions. The algorithm receives a feedback and adjusts its parameter.

We will now solve this problem using the SemiBwK framework. The atoms are the auction and price pair. Formally, $\mathcal{A} := \{(i, p) : i \in [r], p \in S\}$ denote the set of atoms. Additionally, define $\mathcal{A}_i := A \cap \{(i, p) : p \in S\}$. The resource are the underlying items that are being auctioned. For a chosen atom $a_i \in \mathcal{A}_i$, the reward at time-step t is a 0-1 random variable. The realization is 1 if and only if the set of items being auctioned was sold. Similarly, the realized consumption of the resources is 1 for all those items that were sold. Note that, in a time-step the algorithm can choose at most one arm from each of the $\mathcal{A}_i$. Hence, this forms a partition matroid. We will now derive the regret using the SemiBwK-UCB algorithm. Note that $\mathrm{OPT} \leq dB$ and number of atoms is $r|S|$. Hence, regret is $\tilde{O}(d\sqrt{r|S|B})$.

Note that, in general it is unclear how to obtain *low regret* using the BwK framework. One way is to have every subset of parameters for each auction as the set of arms. This makes the number of arms $O(|S|^r)$ and even for moderate values of r the regret is too large. One might try running r seperate instances of BwK, but that would result in budget being violated since the items are *shared* across the auctions and it is unclear a priori how much of each item will be sold in each auction.

Repeated Bidding. This problem is the "flipped" version of the previous one, where the algorithm is the bidder in the auction rather than the auction maker. Consider a bidder who is placing bids in r different repeated auctions. The only resource here is the total money the bidder has. Let us assume that they have a total of B money to place. At each time-step, the bidder has to choose bids $(b_t(1), b_t(2), \ldots, b_t(r))$ to place in the r auctions. For simplicity we will assume that each of the bids $b_t(i)$ are chosen from a finite subset $S \subset [0, 1]$. After placing the bids, based on the

environment of the auctions an outcome vector of tuples is realized. In particular, the outcome vector is $((p_t(1), u_t(1)), (p_t(2), u_t(2)), \ldots, (p_t(r), u_t(r)))$ where $p_t(i)$ denotes the payment the bidder has to make at time t for auction i and $u_t(i)$ is the corresponding utility they receive in auction i. We assume that each of the $p_t(i)$ and $u_t(i)$ lie in $[0, 1]$ and this outcome vector is drawn from a fixed but unknown distribution.

We will now frame this in the `SemiBwK` framework. The atoms correspond to the auction bid pair. In other words, $\mathcal{A} = \{(i, b) : i \in [r], b \in S\}$. There is exactly one resource, which is money and the total budget is B. As before, define $\mathcal{A}_i = \mathcal{A} \cap \{(i, b) : b \in S\}$. An action corresponds to choosing exactly one atom from each of the $\mathcal{A}_i$. This again corresponds to a partition matroid. The reward for a particular action A_t is $\sum_{a \in A_t} u_t(a)$ while the resource consumption is $\sum_{a \in A} p_t(a)$. Number of atoms is $r|S|$ and let U be a known upper bound on OPT $\leq U$. Hence, applying SemiBwK-UCB we get a S-regret of $\tilde{O}(U\sqrt{r|S|/B})$.

7. Proofs

7.1. Proofs from Preliminaries

Proof of Theorem 4.6. (Some minor edits in this proof done to achieve the new Theorem.)

Let S be an arbitrary subset of $\mathcal{Z}_T$. Let $S = S_1 \cup S_2 \cup \ldots S_T$ such that $\forall t \in [T] \quad S_t = \{\zeta_t(a) \in \mathcal{Z}_T \cap S\}$. Define $H_t = \prod_{a \in S_t} \zeta_t(a)$, $G_\tau = \prod_{t \in [\tau]} H_t$, $k_\tau = \sum_{t \in [\tau]} |S_t|$. We will now prove the following equation.

$$\mathbb{E}[G_\tau] \leq 2^{-k_\tau} \tag{7.1}$$

We will prove this by induction on τ.

Base case is when $\tau = 1$. Note that G_τ is just the product of elements in set ζ_1 and they are negatively correlated from the premise. Therefore we are done.

Now for the inductive case of $\tau \geq 2$,

$$\begin{aligned} \mathbb{E}[H_\tau | \mathcal{Z}_{\tau-1}] &\leq \prod_{a \in S_\tau} \mathbb{E}[\zeta_\tau(a) | \mathcal{Z}_{\tau-1}] && \text{From negative correlation} && (7.2) \\ &\leq 2^{-|S_\tau|} && \text{From assumption in Lemma 4.6} && (7.3) \end{aligned}$$

Therefore, we have

$$\begin{aligned} \mathbb{E}[G_\tau] &= \mathbb{E}[\mathbb{E}[G_{\tau-1} H_\tau | \mathcal{Z}_{\tau-1}]] && \text{Law of iterated expectation} \\ &= \mathbb{E}[G_{\tau-1} \mathbb{E}[H_\tau | \mathcal{Z}_{\tau-1}]] && \text{Since } G_{\tau-1} \text{ is a fixed value conditional on } \mathcal{Z}_{\tau-1} \\ &\leq 2^{-|S_\tau|} \mathbb{E}[G_{\tau-1}] && \text{From Equation 7.3} \\ &\leq 2^{-k_\tau} && \text{From inductive hypothesis} \end{aligned}$$

Therefore, we have that $\mathcal{Z}_T$ satisfy Equation 2.1.

A symmetric argument follows, by considering the family of random variables $\mathcal{Z}'_T = \{1 - \zeta_t(a) : a \in [n], t \in [T]\}$. Hence, they are negatively correlated.

7.2. Proofs from Section 4.1 (analysis of LPs)

Proof of Claim 4.4. Note, we only need to prove the first inequality. We will prove it as follows.

Let $\boldsymbol{x}^*$ denote an optimal solution to $\text{LP}_{\texttt{BwK}}$(B). Consider $(1-\epsilon)x^*$; this is feasible to $\text{LP}_{\texttt{BwK}}(B_\epsilon)$, since for every S, $(1-\epsilon)x^*(S) \leq 1$ and $\sum_{S\subseteq[n]:S\in\mathcal{S}} C(S,j)(1-\epsilon)x(S) \leq B_\epsilon/T$. Hence, this is a feasible solution. Now, consider the objective function. Let $\boldsymbol{y}$ denote an optimal solution to $\text{LP}_{\texttt{BwK}}(B_\epsilon)$. We have that

$$\text{OPT}_{\texttt{BwK}}(B_\epsilon) = \sum_{S\subseteq[n]:S\in\mathcal{S}} \mu(S)y^*(S) \geq \sum_{S\subseteq[n]:S\in\mathcal{S}} \mu(S)(1-\epsilon)x^*(S) = (1-\epsilon)\,\text{OPT}_{\texttt{BwK}}(B)$$

Proof of Claim 4.5. We will first prove the second inequality.

Consider the optimal solution vector $\boldsymbol{x}$ to $\text{LP}_{\texttt{BwK}}(B_\epsilon)$. Define $S^* := \{S : x(S) \neq 0\}$.

We will now map this to a feasible solution to LP_{ATOMS} and show that the objective value does not decrease. This will then complete the claim. Consider the following solution $\boldsymbol{y}$ defined as follows.

$$y(a) = \sum_{S\in S^*:a\in S} x(S)$$

We will now show that $\boldsymbol{y}$ is a feasible solution to the polytope $\mathcal{P}$. From the definition of $\boldsymbol{y}$, we can write it as $\boldsymbol{y} = \sum_{S\in S^*} x(S) \times \mathbb{I}[S]$. Here, $\mathbf{I}[S]$ is a binary vector, such that it has 1 at position a if and only if atom a is present in set S. Hence, $\boldsymbol{y}$ is a point in the polytope since it can be written as convex combination of its vertices.

Now, we will show that, $\boldsymbol{y}$ also satisfies the resource consumption constraint.

$$\boldsymbol{C(j)}\cdot\boldsymbol{y} = \sum_{a=1}^{n} C(a,j) \sum_{S\in S^*:a\in S} x(S) = \sum_{S\in S^*}\sum_{a\in S} C(a,j)x(S) = \sum_{S\in S^*} C(S,j)x(S) \leq B_\epsilon/T$$

The last inequality is because in the optimal solution, the x value corresponding to subset S^* is 1 while rest all are 0. We will now show that $\boldsymbol{y}$ produces an objective value at least as large as $\boldsymbol{x}$.

$$\begin{aligned}\text{OPT}_{\texttt{atoms}} = \boldsymbol{\mu}\cdot\boldsymbol{y}^* \geq \boldsymbol{\mu}\cdot\boldsymbol{y} &= \sum_{a=1}^{n}\mu(a)\sum_{S\in S^*:a\in S} x(S)\\ &= \sum_{S\in S^*}\sum_{a\in S}\mu(a)x(S) = \sum_{S\in S^*}\mu(S)x(S) = \text{OPT}_{\texttt{subsets}}(B_\epsilon)\end{aligned}$$

Now we will prove the first inequality.

Consider a time t. Given an optimal solution $\boldsymbol{x}^*$ to LP_{ATOMS} we will show that this is feasible to $\text{LP}_{\text{ALG},t}$. Note that, $\boldsymbol{x}^*$ satisfies the constraint set $\boldsymbol{x}\in\mathcal{P}$ since that is same for both $\text{LP}_{\text{ALG},t}$ and LP_{ATOMS}. Now consider the constraint $\boldsymbol{C_t^-(j)}\cdot\boldsymbol{x} \leq \frac{B_\epsilon}{T}$. Note that $C_t^-(a,j) \leq C(a,j)$. Hence, we have that $\boldsymbol{C_t^-(j)}\cdot\boldsymbol{x}^* \leq \boldsymbol{C(j)}\cdot\boldsymbol{x}^* \leq \frac{B_\epsilon}{T}$. The last inequality is because $\boldsymbol{x}^*$ is a feasible solution to LP_{ATOMS}.

Now consider the objective function. Let $\boldsymbol{y}^*$ denote the optimal solution to $\text{LP}_{\text{ALG},t}$.

$\text{OPT}_{\texttt{ALG},t} = \boldsymbol{\mu_t^+}\cdot\boldsymbol{y}^* \geq \boldsymbol{\mu_t^+}\cdot\boldsymbol{x}^* \geq \boldsymbol{\mu}\cdot\boldsymbol{y}^* = \text{OPT}_{\texttt{atoms}}$.

7.3. Proofs for the "rewards" clean event Lemma 4.7

Proof of Equation 4.2. Recall that $r_t = \boldsymbol{\mu_t Y_t}$. Note that, $\mathbb{E}[\boldsymbol{\mu_t Y_t}] = \boldsymbol{\mu Y_t}$ when the expectation is taken over just the independent samples of μ. Hence, from Theorem A.2 we have with probability $1 - \exp(-\Omega(\alpha))$

$$\begin{aligned}
|\sum_{t\le T} r_t - \sum_{t\le T} \boldsymbol{\mu}\cdot\boldsymbol{Y_t}| &\le 3nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\le T}\boldsymbol{\mu}\cdot\boldsymbol{Y_t}\ ,\ nT\right)\\
&\le 3nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\le T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{Y_t}\ ,\ nT\right)\\
&\le 3nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\le T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{x_t}\ ,\ nT\right)
\end{aligned}$$

The last inequality is because Y_t is a feasible solution to $\mathrm{LP}_{\mathrm{ALG}}$.

Proof of Equation 4.3. For this part, the arguments similar to Agrawal and Devanur (2014b) follow with some minor adaptations. For sake of completeness we describe the full proof. Note that we have,

$$|\sum_{t\le T}\boldsymbol{\mu}\cdot\boldsymbol{Y_t} - \boldsymbol{\mu_t^+}\cdot\boldsymbol{Y_t}| \le \sum_{a=1}^{n}|\sum_{t\le T}\mu(a)Y_t(a) - \mu_t^+(a)Y_t(a)|$$

Now, using Lemma A.1 in Appendix, we have that with probability $1 - nT\exp(-\Omega(\alpha))$

$$|\sum_{t\le T}\mu(a)Y_t(a) - \mu_t^+(a)Y_t(a)| \le 12\sum_{t\le T}\mathrm{Rad}(\mu(a), N_t(a)+1)$$

Hence, we have

$$\begin{aligned}
\sum_{a=1}^{n}|\sum_{t\le T}\mu(a)Y_t(a) - \mu_t^+(a)Y_t(a)| &= 12\sum_{a=1}^{n}\sum_{r=1}^{N_T(a)+1}\mathrm{Rad}(\mu(a), r)\\
&\le 12\sum_{a=1}^{n}(N_T(a)+1)\,\mathrm{Rad}(\mu(a), N_T(a)+1)\\
&\le 12\sqrt{\gamma n\,(\boldsymbol{\mu}\cdot(\boldsymbol{N_T}+\mathbf{1}))} + 12\gamma n
\end{aligned}$$

The last inequality is from the definition of Rad function and using the Cauchy-Swartz inequality. Note that $\boldsymbol{\mu N_T} = \sum_{t\le T}\boldsymbol{\mu}\cdot\boldsymbol{Y_t}$. Also, since we have $\mu(a) \le \mu_t^+(a)$, we have

$$12\sqrt{\gamma n\,(\boldsymbol{\mu}\cdot(\boldsymbol{N_T}+\mathbf{1}))} + 12\gamma n \le 12\sqrt{\gamma n\left(\sum_{t\le T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{Y_t}\right)} + 12\sqrt{\gamma}n + 12\gamma n$$

Finally note that $\boldsymbol{Y_t}$ is a feasible solution to the semi-bandit polytope $\mathcal{P}$. Hence, we have that

$$\boldsymbol{\mu_t^+} \cdot \boldsymbol{Y_t} \leq \boldsymbol{\mu} \cdot \boldsymbol{x_t}$$

Hence,

$$12\sqrt{\gamma n\left(\sum_{t\leq T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{Y_t}\right)}+12\sqrt{\gamma}n+12\gamma n\leq 12\sqrt{\gamma n\left(\sum_{t\leq T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{x_t}\right)}+12\sqrt{\gamma}n+12\gamma n$$

Proof of Lemma 4.7. Denote $H=\sqrt{\sum_{t\leq T}\boldsymbol{\mu_t^+}\cdot\boldsymbol{x_t}}$. From 4.2, 4.3 and 4.4, we have that $H^2-2\Omega(\sqrt{\gamma n})H\leq\sum_{t\leq T}r_t+O(\gamma n)$. Hence, re-arranging this gives us $H\leq\sqrt{\sum_{t\leq T}r_t}+O(\sqrt{\gamma n})$. Now plugging this back into equations 4.2, 4.3 and 4.4 proves the Lemma 4.7.

7.4. Proof for "Consumption" Clean Event Lemma 4.8

As we did for clean event described by Lemma 4.7, we will split the proof into following three equations. Fix an arbitrary resource $j\in[d]$. With probability at least $1-nT\exp(-\Omega(\alpha))$ the following holds:

$$|\sum_{t\leq T}\chi_t(j)-\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}|\leq 3nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}\,,\,nT\right)\tag{7.4}$$

$$|\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}-\boldsymbol{C_t^-(j)}\cdot\boldsymbol{Y_t}|\leq 12\sqrt{\gamma n\left(\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}\right)}+12\sqrt{\gamma}n+12\gamma n\tag{7.5}$$

$$|\sum_{t\leq T}\boldsymbol{C_t^-(j)}\cdot\boldsymbol{Y_t}-\boldsymbol{C_t^-(j)}\cdot\boldsymbol{x_t}|\leq nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}\,,\,nT\right)\tag{7.6}$$

Using the parts 7.4 and 7.6 we can now find the following upper bound on $\sqrt{\sum_{t\leq T}\boldsymbol{C_t(j)}\cdot\boldsymbol{Y_t}}$. Hence, combining Lemmas 7.4, 7.5 and 7.6 with the above bound and taking an Union Bound over all the resources, we get Lemma 4.8.

Proof of Equation 7.4. We have that $\{C_t(a,j):a\in[n]\}$ is a set of independent random variables over a probability spacee C_Ω. Note that, $\mathbb{E}_{C_\Omega}C_t(a,j)Y_t(a)=C(a,j)Y_t(a)$. Hence, we can invoke Theorem A.2 on *independent random variables* to get with probability $1-\exp(-\Omega(\alpha))$

$$|\sum_{t\leq T}\chi_t(j)-\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}|\leq 3nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\leq T}\boldsymbol{C(j)}\cdot\boldsymbol{Y_t}\,,\,nT\right)$$

Proof of Equation 7.5. This is very similar to proof of 4.3 and we will skip the repetitive parts. Hence, we have with probability $1-nT\exp(-\Omega(\alpha))$

$$|\sum_{t\le T} \boldsymbol{C(j)\cdot Y_t} - \boldsymbol{C_t^-(j)\cdot Y_t}| \le 12\sqrt{\gamma n(\boldsymbol{C(j)\cdot(N_T+1)})} + 12\gamma n$$

$$\le 12\sqrt{\gamma n\left(\sum_{t\le T}\boldsymbol{C(j)\cdot Y_t}\right)} + 12\sqrt{\gamma}n + 12\gamma n$$

Proof of Equation 7.6. Recall that $C_t^-(a,j)$ is determined by the random variables $\mathcal{F}_{t-1} := \{Y_{t'}(a') : \forall t' < t, \forall a \in [n]\}$. Following the arguments for rewards case, we have that conditioned on $\mathcal{F}_{t-1}$, the random variables $\{Y_t(a) : a \in [n]\}$ to be negatively correlated. Define $\tilde{\zeta}_t(a) = C_t^-(a)Y_t(a)$ for all $t \in [T], a \in [n]$. Once again note that, $\mathbb{E}[C_t^-(a)Y_t(a) \mid \mathcal{F}_{t-1}] = C_t^-(a)x_t(a)$. Applying Theorem 4.6 with $\zeta_t(a) = (1 + \tilde{\zeta}_t(a) - C_t^-(a)x_t(a))/2$ for all $t \in [T], a \in [n]$, we have that the family $\{\zeta_t(a) : t \in [T], a \in [n]\}$ is negatively correlated. Hence, applying the Chernoff bounds from Theorem 2.3 to this family, we obtain $\Pr[\frac{1}{nT}(\sum_{t=1}^{T}\sum_{a=1}^{n}\tilde{\zeta}_t(a) - C_t^-(a)x_t(a)) \ge \epsilon] \le c.e^{-nT\mathcal{D}_{\mathrm{KL}}(1/2+\epsilon\,\|\,1/2)}$. And using the standard transformation from the additive form to the multiplicative form, we obtain Theorem 2.2 for the family $\{\tilde{\zeta}_t(a) : t \in [T], a \in [n]\}$ which implies Theorem 2.4. Hence, we have with probability $1 - \Omega(\exp(-\alpha))$

$$|\sum_{t\le T}\boldsymbol{C_t^-(j)\cdot Y_t} - \boldsymbol{C_t^-(j)\cdot x_t}| \le nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\le T}\boldsymbol{C_t^-(j)\cdot Y_t}\,,\; nT\right)$$

$$\le nT\,\mathrm{Rad}\left(\frac{1}{nT}\sum_{t\le T}\boldsymbol{C(j)\cdot Y_t}\,,\; nT\right)$$

Proof of Lemma 4.8. Denote $G = \sqrt{\sum_{t\le T}\boldsymbol{C(j)\cdot Y_t}}$. From Equation 7.4, 7.5 and 7.6, we have that $G^2 - 2\Omega(\sqrt{\gamma n})G \le \sum_{t\le T}\boldsymbol{C_t^-(j)\cdot x_t} + O(\gamma n)$. Note that $\sum_{t\le T}\boldsymbol{C_t^-(j)\cdot x_t} \le B_\epsilon$. Hence, $G^2 - 2\Omega(\sqrt{\gamma n})G \le B_\epsilon + O(\gamma n)$. Hence, re-arranging this gives us $G \le \sqrt{B_\epsilon} + O(\sqrt{\gamma n})$. Plugging this back in Equations 7.4, 7.5 and 7.6, we get Lemma 4.8.

7.5. Combining the clean events to obtain regret bound in (4.1)

Recall that from Claim 4.3, we have $\mathrm{OPT}_{\texttt{ALG}_\gamma} \ge \frac{1}{T}(1-\epsilon)\,\mathrm{OPT}$. Let us define the performance of the algorithm as $\mathrm{ALG} = \sum_{t\le T} r_t$. From Lemma 4.7, we have that with probability $1 - ndT\exp(-\Omega(\alpha))$

$$\begin{aligned}\mathrm{ALG} &\ge (1-\epsilon)\,\mathrm{OPT} - O(\sqrt{\gamma n\,\mathrm{ALG}}) - O(\gamma n)\\ &\ge (1-\epsilon)\,\mathrm{OPT} - O(\sqrt{\gamma n\,\mathrm{OPT}}) - O(\gamma n) \qquad \text{Since, } \mathrm{ALG} \le \mathrm{OPT}\end{aligned}$$

Choosing $\epsilon = \sqrt{\frac{\gamma n}{B}} + \frac{\gamma n}{B}$, we have Equation 4.1. Additionally note that, for a given δ, we have to set $\gamma = \Omega(\log(\frac{ndT}{\delta}))$.

Now we will argue that the algorithm doesn't exhaust the resource budget before time-step T with probability $1 - ndT\exp(-\Omega(\alpha))$. Note that for every resource $j \in [d]$, $\sum_{t\leq T} \boldsymbol{C_t^-(j)} \cdot \boldsymbol{x_t} \leq (1-\epsilon)B$. Hence, combining this with Claim 4.8, we have $\sum_{t\leq T} C_t(j) \leq (1-\epsilon)B + \epsilon B \leq B$

7.6. Proof Sketch for extension (Section 5)

Here, we will give a sketch for proving the regret in (5.1). Note that, the RRS guarantees no form of concenteration among the atoms. Hence, we analyze it atom-by-atom and take an union bound across all the atoms. For the rewards clean event, we obtain with probability $1 - n^2T\exp(-\Omega(\alpha))$

$$|\sum_{t\leq T} r_t - \sum_{t\leq T} \boldsymbol{\mu_t^+} \cdot \boldsymbol{x_t}| \leq \sum_{a=1}^{n} O(\sqrt{\gamma n \sum_{t\leq T} r_t(a)}) + O(\gamma n^2)$$

Translating this to the final regret calculation, we obtain

$$OPT - ALG \leq O\left(\mathrm{OPT}\sqrt{\frac{\gamma n}{B}} + \sum_{a=1}^{n}\sqrt{\gamma n r_t(a)} + \gamma n^2\right)$$

Now, on the RHS we want to maximize $\sum_{a=1}^{n}\sqrt{r_t(a)}$ subject to $\sum_{a=1}^{n} r_t(a) = \mathrm{OPT}$. Using a standard Lagrangian calculation, we obtain that the maximizer is when $r_t(a) = \frac{\mathrm{OPT}}{n}$ for all the atoms a. Hence, we have the regret in (5.1).

Now, let us look at the resources. Here, we will critically use the fact that each atom has a dedicated resource. Since, every atom has a dedicated resource, while analyzing a particular resource, we need to consider just the corresponding arm. In other words, we have

$$\forall j \in [d] \quad |\sum_{t\leq T} \chi_t(j) - \sum_{t\leq T} \boldsymbol{C_t^-(j)} \cdot \boldsymbol{x_t}| = |\sum_{t\leq T} C_t(a_j, j).Y_t(a_j) - \sum_{t\leq T} C_t^-(a_j, j)x_t(a_j)|$$

Here, a_j is the arm dedicated to resource j. Since $\{Y_t(a_j) : t \leq T\}$ form a Martingale, we can use the concenteration bounds similar to sampling a single atom and the arguments in Agrawal and Devanur (2014b) goes as-is. Hence, with probability $1 - ndT\exp(-\Omega(\alpha))$ the algorithm does not run out of resources before time-step T.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *25th Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *15th ACM Conf. on Economics and Computation (ACM EC)*, 2014a.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006. ACM, 2014b.

Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *29th Advances in Neural Information Processing Systems (NIPS)*, 2016.

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *29th Conf. on Learning Theory (COLT)*, 2016.

Arash Asadpour, Michel X Goemans, Aleksander Madry, Shayan Oveis Gharan, and Amin Saberi. An o (log n/log log n)-approximation algorithm for the asymmetric traveling salesman problem. In *SODA*, volume 10, pages 379–389. SIAM, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Moshe Babaioff, Shaddin Dughmi, Robert D. Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Trans. on Economics and Computation*, 3(1):4, 2015. Special issue for *13th ACM EC*, 2012.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Yaron Singer. Learning on a budget: posted price mechanisms for online procurement. In *13th ACM Conf. on Electronic Commerce (EC)*, pages 128–145, 2012.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *54th IEEE Symp. on Foundations of Computer Science (FOCS)*, 2013.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *27th Conf. on Learning Theory (COLT)*, 2014.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57:1407–1420, 2009.

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

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1), 2012.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1190–1204. Society for Industrial and Applied Mathematics, 2013.

Chandra Chekuri, Jan Vondrak, and Rico Zenklusen. Dependent randomized rounding via exchange properties of combinatorial structures. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 575–584. IEEE, 2010.

Chandra Chekuri, Jan Vondrák, and Rico Zenklusen. Multi-budgeted matchings and matroid intersection via dependent rounding. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1080–1097. SIAM, 2011.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 151–159. JMLR Workshop and Conference Proceedings, 2013.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and marc lelarge. Combinatorial bandits revisited. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2116–2124. Curran Associates, Inc., 2015.

Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson, and S. Muthukrishnan. Adaptive submodular maximization in bandit setting. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2697–2705. 2013.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations, October 2012.

Rajiv Gandhi, Samir Khuller, Srinivasan Parthasarathy, and Aravind Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

Sudipta Guha and Kamesh Munagala. Multi-armed Bandits with Metric Switching Costs. In *36th Intl. Colloquium on Automata, Languages and Programming (ICALP)*, pages 496–507, 2007.

Anupam Gupta, Ravishankar Krishnaswamy, Marco Molinaro, and R. Ravi. Approximation algorithms for correlated knapsacks and non-martingale bandits. In *52nd IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 827–836, 2011.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *J. of Machine Learning Research (JMLR)*, 8:2369–2403, 2007.

Russell Impagliazzo and Valentine Kabanets. Constructive proofs of concentration bounds. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 617–631. Springer, 2010.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvári, and Zheng Wen. DCM bandits: Learning to rank with multiple clicks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1215–1224, 2016.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. Working paper, published at `http://arxiv.org/abs/1312.1277`, 2015. Merged and revised version of conference papers in *ACM STOC 2008* and *ACM-SIAM SODA 2010*.

Akshay Krishnamurthy, Alekh Agarwal, and Miroslav Dudík. Contextual semibandits via supervised learning oracles. In *29th Advances in Neural Information Processing Systems (NIPS)*, 2016.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In Nevin L. Zhang and Jin Tian, editors, *UAI*, pages 420–429. AUAI Press, 2014.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 767–776. JMLR Workshop and Conference Proceedings, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvri. Tight regret bounds for stochastic combinatorial semi-bandits. In Guy Lebanon and S. V. N. Vishwanathan, editors, *AISTATS*, JMLR Workshop and Conference Proceedings. JMLR.org, 2015b.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Alessandro Panconesi and Aravind Srinivasan. Fast randomized algorithms for distributed edge coloring. In *Proceedings of the Eleventh Annual ACM Symposium on Principles of Distributed Computing*, PODC '92, pages 251–262, 1992.

Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1982.

Prabhakar Raghavan and Clark D Tompson. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374, 1987.

Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2002.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd Intl. World Wide Web Conf. (WWW)*, pages 1167–1178, 2013.

Aleksandrs Slivkins. Dynamic ad allocation: Bandits with budgets. A technical report on `arxiv.org/abs/1306.0155`, June 2013.

Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *SIGecom Exchanges*, 12(2), December 2013.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285294, 1933.

Long Tran-Thanh, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. ϵ-first policies for budget-limited multi-armed bandits. In *24th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1211–1216, 2010.

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *26th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1134–1140, 2012.

Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In Francis R. Bach and David M. Blei, editors, *ICML*, JMLR Workshop and Conference Proceedings, pages 1113–1122. JMLR.org, 2015.

David P Williamson and David B Shmoys. *The design of approximation algorithms*. Cambridge university press, 2011.

Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton. Cascading bandits for large-scale recommendation problems. 2016.

Appendix A. Technical Lemmas and Proofs

In this section, we will list some of the technical lemmas we used in the main section of the paper.

Lemma A.1 (Agrawal and Devanur (2014b)) *Let $\hat{\mu}_t(a)$ denote the empirical average of the mean reward at time t for atom a. Similarly, let $\hat{C}_t(a,j)$ denote the empirical average of the mean consumption of resource j for atoms a at time t. Let $N_t(a)$ be the number of times atom a has been played till beginning of time t. Then with probability at least $1-\exp[-\Omega(\gamma)]$, we have the following*

$$|\hat{\mu}_t(a) - \mu_t(a)| \leq 2\,\mathrm{Rad}(\hat{\mu}_t(a), N_t(a)+1) \tag{A.1}$$

$$\forall j \in [d] \quad |\hat{C}_t(a,j) - C_t(a,j)| \leq 2\,\mathrm{Rad}(\hat{C}_t(a,j), N_t(a)+1) \tag{A.2}$$

Theorem A.2 (Babaioff et al. (2015)) *Let $X_1, X_2, \ldots, X_m$ denote a set of random variables. For each t, let α_t denote the multiplier determined by random variables $X_1, X_2, \ldots, X_{t-1}$. Let $M = \sum_{t=1}^{\tau} M_t$ where $M_t = \mathbb{E}[X_t | X_1, X_2, \ldots, X_{t-1}]$. Then for any $b \geq 1$, we have the following with probability at least $1 - m^{\Omega(-b)}$:*

$$|\sum_{t=1}^{\tau} \alpha_t (X_t - M_t)| \leq b(\sqrt{M \log m} + \log m)$$

Proof of Theorem 2.4 The proof of this follows very similar to that by Kleinberg et al. (2015) and others. We will split it into two cases.

- $\mu \geq \frac{\alpha}{6g}$: Invoking Theorem 2.2, Equation (a) with $\epsilon = \frac{1}{2}\sqrt{\frac{\alpha}{6g\mu}}$ we have that with probability $1 - \exp[-\alpha/18]$ that $|X - \mu| \leq \epsilon\mu$. Note that $\epsilon\mu \leq \frac{\mu}{2}$. Hence,

$$
\begin{aligned}
|X-\mu| &< \frac{1}{2}\sqrt{\frac{\alpha\mu}{6g}} \\
&\leq \frac{1}{2}\sqrt{\frac{2\alpha X}{6g}} && \text{Since } |X-\mu| \leq \mu/2 \\
&\leq 1.5\,\mathrm{Rad}_\alpha(X,g)
\end{aligned}
$$

- $\mu \leq \frac{\alpha}{6g}$: Invoke Theorem 2.2 Equation (b) with $a = \frac{\alpha}{g}$. Then with probability $1 - 2^{-\alpha}$ we have $X < \frac{\alpha}{g}$. Therefore,

$$
|X-\mu| \leq \frac{\alpha}{g} \leq \mathrm{Rad}_\alpha(X,g) \leq 3\,\mathrm{Rad}_\alpha(\mu,g)
$$

Finally, in the above argument replacing X with Z and μ with $\hat{\mu}$ and using the versions in Eq. 2.5 and Eq. 2.6 instead of Theorem 2.2 will give the second part of Theorem 2.4

Proof of Claim 2.1

Minor edits to handle the second property of negative correlation.

Define $Y_i := 1 + X_i - \mu_i$ for all $i \in [m]$. Define $Z_i := 1 - \mu_i$ for all $i \in [m]$. Note that $Y_i = X_i + Z_i$ and $Z_i \geq 0, X_i \geq 0$.

Consider a subset $S \subseteq [m]$. We have,

$$
\begin{aligned}
\mathbb{E}[\prod_{i\in S} Y_i] &= \mathbb{E}[\sum_{T\subseteq S}\prod_{i\in T} X_i \prod_{j\in S\setminus T} Z_i] \\
&= \sum_{T\subseteq S}\mathbb{E}[\prod_{i\in T} X_i]\prod_{j\in S\setminus T} Z_i && \text{From Linearity of Expectation} \\
&\leq \sum_{T\subseteq S}\prod_{i\in T}\mu_i \prod_{j\in S\setminus T} Z_i && \text{fact that } Z_i, X_i \geq 0 \text{ and } X_i \text{ are negatively correlated} \\
&= \prod_{i\in S}(1-\mu_i+\mu_i) = 1 && \text{From Binomial Theorem}
\end{aligned}
$$

Note that neg. correlation property is preserved under scaling by a positive constant factor.

Consider $Y_i' := 1 - (1 + X_i - \mu_i)/2$ for all $i \in [m]$. Note that, we have $Y_i' = (1 + \mu_i - X_i)/2$ for all $i \in [m]$.. Hence, setting $Z_i' := 1 - X_i$ for all $i \in [m]$ and noting that $Y_i' = \mu_i + Z_i'$ for $\mu_i \geq 0, Z_i' \geq 0$, a similar argument as above follows.

Proof for (2.6) implied from (2.5) This is a standard argument and here we borrow this from proof of Theorem 4.4 in Mitzenmacher and Upfal (2005).

Let $a = (1+\epsilon)\hat{\mu}$ in the first equation of (2.5). Since, $a \geq 6\hat{\mu}$, we have that $\epsilon = a/\hat{\mu} - 1 \geq 5$. Hence,

$$\begin{aligned}\Pr[Z \geq (1+\epsilon)\hat{\mu}] &\leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{nT\hat{\mu}} \\ &\leq \left(\frac{e}{1+\epsilon}\right)^{(1+\epsilon)nT\hat{\mu}} \\ &\leq \left(\frac{e}{6}\right)^{anT}) \\ &\leq 2^{-anT}\end{aligned}$$

Appendix B. Formal Definition for Various special cases of Semi-Bandit Constraints

In this section, we will give formal definition of various families of subsets inducing the semi-bandit constraint, for which our algorithms work. Additionally, we consider the *relaxation*; this means the constraint $x(a) \in \{0,1\}$ is relaxed to $x(a) \in [0,1]$.

Cardinality Constraint
This is the simplest semi-bandit constraint our algorithm is applicable to. In this constraint, the algorithm is allowed to choose any subset of atoms such that the number of atoms in the subset is at most a fixed value K. We can define this polytope our algorithm uses for this constraint as follows:

$$\begin{aligned}&\sum_{a\in n} x(a) \leq K \\ &1 \geq x(a) \geq 0 \quad \forall a \in [n]\end{aligned} \tag{LP-Cardinality}$$

Spanning Tree
In this case we are given a graph $G = (V, E)$. The atoms correspond to the edges of this graph. In each step the algorithm can choose a subset of edges such that they form a spanning tree to the graph G. A spanning tree is a subset of edges such that, in the sub-graph induced by this subset for any two vertices s, t, there is a path from s to t. We will denote $E(S)$ to denote the edges in the sub-graph induced by S and $x(E(S)) := \sum_{e\in E(S)} x_e$. Formally, this semi-bandit constraint can be represented by the following polytope:

$$\begin{aligned}&x(E(S)) \leq |S| - 1 \quad \forall S \subseteq V \\ &x(E(V)) = |V| - 1 \\ &1 \geq x_e \geq 0 \quad \forall e \in E\end{aligned} \tag{LP-SpanningTree}$$

Matroids
A matroid $M(E, \mathcal{I})$ is defined by a set of ground atoms E and a set of subsets of atoms in E called *independent sets* which is denoted by $\mathcal{I}$. Additionally, $M(E, \mathcal{I})$ satisfies a few properties.

- **Empty Set**: The empty set ϕ is present in $\mathcal{I}$
- **Hereditary property**: For two subsets $X, Y \subseteq E$ such that $X \subseteq Y$, we have that

$$Y \in \mathcal{I} \implies X \in \mathcal{I}$$

- **Exchange property**: For $X, Y \in \mathcal{I}$ and $|X| > |Y|$, we have that

$$\exists e \in X \setminus Y : Y \cup \{e\} \in \mathcal{I}$$

Matroid polytope: We will now describe a characterization of the matroid polytopes via linear constraints. The proof of correctness is well-studied and the reader is encouraged to refer to a standard text-book in combinatorial optimization for more details.

For a matroid M, define $r(M) : 2^E \to \mathbb{N}$ to be the rank function defined as

$$r(M) = \max\{|Y| : Y \subseteq X, Y \in \mathcal{I}\}$$

The matroid polytope then can be defined via the following polytope. Here, x_e for every atom e of E is the set of variables. Additionally, the following definition is used - $x(S) := \sum_{e \in S} x_e$

$$\begin{array}{ll} x(S) \leq r(S) & \forall S \subseteq E \\ x_e \geq 0 & \forall e \in E \\ \boldsymbol{x} \in \mathbb{R}^E & \end{array} \qquad \text{(LP-Matroid)}$$

Partition Matroid: In our applications, we will encouter this well-known matroid called the partition matroid. Suppose we have some n ground elements. We have a collection B_i of disjoint subsets of $[n]$, and real numbers $0 \leq d_i \leq |B_i|$. A set I is independent if and only if for every i, $|I \cap B_i| \leq d_i$. This system of ground elements and independent sets is called a *partition matroid*.