Covariance-adapting algorithm for semi-bandits with application to sparse outcomes

Pierre Perrault, Vianney Perchet, Michal Valko

Proceedings of Machine Learning Research vol 125:1–33, 2020 33rd Annual Conference on Learning Theory

Covariance-adapting algorithm for semi-bandits with application to sparse outcomes

Pierre Perrault PIERRE.PERRAULT@OUTLOOK.COM
Inria Lille & ENS Paris-Saclay

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
ENSAE & Criteo AI Lab

Michal Valko VALKOM@DEEPMIND.COM
DeepMind Paris

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We investigate *stochastic combinatorial semi-bandits*, where the entire joint distribution of outcomes impacts the complexity of the problem instance (unlike in the standard bandits). Typical distributions considered depend on specific parameter values, whose prior knowledge is required in theory but quite difficult to estimate in practice; an example is the commonly assumed *sub-Gaussian* family. We alleviate this issue by instead considering a new general family of *sub-exponential* distributions, which contains bounded and Gaussian ones. We prove a new lower bound on the regret on this family, that is parameterized by the *unknown* covariance matrix, a tighter quantity than the sub-Gaussian matrix. We then construct an algorithm that uses covariance estimates, and provide a tight asymptotic analysis of the regret. Finally, we apply and extend our results to the family of sparse outcomes, which has applications in many recommender systems.

Keywords: combinatorial stochastic semi-bandits, covariance, sparsity, confidence ellipsoid

1. Introduction

Complete automatic adaptation of algorithms to the processed data, as opposed to the requirement of prior knowledge on underlying structure or to some manual tuning of parameters, is one of the fundamental challenges in machine learning. We address this challenge for *stochastic (combinatorial) semi-bandits*, and provide an algorithm adaptive to the correlation structure of the data, leading to provably faster learning in a sequential setting with limited feedback.

Stochastic multi-arm bandits (MAB) are decision-making problems where an *agent* sequentially acts in an uncertain environment. At each round $t \in \mathbb{N}^*$, the agent selects an arm i among a ground set $[n] \triangleq \{1, \dots, n\}$ of $n \in \mathbb{N}^*$ arms. This choice generates some reward (or outcome) $X_{i,t} \in \mathbb{R}$, a random variable drawn from $\mathbb{P}_{X_i}$, independently from previous rounds, where $\mathbb{P}_{X_i}$ is some probability distribution — *unknown* to the agent — of mean μ_i^*. The objective of the agent is to maximize the expected cumulative reward, or equivalently, to minimize the *regret*, defined as the difference between the expected cumulative reward achieved by always selecting the single optimal arm and that achieved by the agent. To accomplish this objective (Robbins, 1952), the agent must trade-off between *exploration* (gaining information about the arm distributions) and *exploitation* (greedily using the information collected so far). To assess the learning policy followed by the agent (also called a *learning algorithm*), upper bounds on the regret are often derived as a guarantee

on its performance. These bounds are valid provided that $(\mathbb{P}_{X_1}, \dots, \mathbb{P}_{X_n})$ belongs to some family of probability distributions, e.g., the family of sub-Gaussian outcomes.

There exist sophisticated learners adaptive to the environment, in the sense that their performance guarantees improve (or stated otherwise, their regret upper bounds decrease) when the problem instance is "simpler" for some appropriate notions of complexity. For instance, Audibert et al. (2009) and Mukherjee et al. (2017) proposed to estimate the variance of each arm to construct adaptive confidence intervals for each mean μ_i^*, based on Bernstein's inequality. This leads to an algorithm having variance-dependent regret bounds. Garivier and Cappé (2011) went beyond variance estimation and proposed a *Kullback–Leibler divergence* based confidence region, and provided a tighter regret upper bound. Thompson sampling can also offer such adaptive regret upper bounds (Kaufmann et al., 2012). Our objective is to attain such adaptivity, but for the challenging combinatorial extension of bandits, called stochastic semi-bandits, described next.

Henceforth, for notation conveniences, we typeset vectors in bold and indicate components with indices, i.e., $\mathbf{a} = (a_i)_{i\in[n]} \in \mathbb{R}^n$. In *combinatorial semi-bandits* (Cesa-Bianchi and Lugosi, 2012), the action space $\mathcal{A}$ is a collection of *subset* of arms. At each round t, the agent chooses some action $A_t \in \mathcal{A}$, receives the *total* reward associated to the selected actions A_t, assumed to be $\sum_{i\in A_t} X_{i,t}$, and observes the outcome of each base arm of A_t, i.e., the vector $(X_{i,t}\mathbb{I}\{i \in A_t\})_{i\in[n]}$. The action space $\mathcal{A}$ depends on the combinatorial problem at hand. For example, actions in $\mathcal{A}$ could be a path from an origin to a destination in a network (György et al., 2007; Talebi et al., 2013) or a subset of items to recommend to a customer (Wang et al., 1997). Many other examples and applications are given by Cesa-Bianchi and Lugosi (2012). Notice that in this setting, the whole joint distribution of the vector of outcomes is relevant, contrary to standard bandit problems where only the n marginals are sufficient to characterize the difficulty of the instance. If we define $\mathbf{X} \triangleq (X_1, \dots, X_n)$, the objective is to design a learning algorithm adaptive to the distribution $\mathbb{P}_{\mathbf{X}}$. This is more challenging than in standard bandits, where adaptivity is only with respect to $\otimes_{i\in[n]}\mathbb{P}_{X_i}$.

In a first approach, Degenne and Perchet (2016) considered the general family of $\mathbf{C}$-sub-Gaussian probability distributions, with $\mathbf{C} \succeq 0$ (i.e., $\mathbf{C}$ is positive semi-definite). Formally, those distributions $\mathbb{P}_{\mathbf{X}}$ of mean $\boldsymbol{\mu}^*$ satisfy

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^n,\ \mathbb{E}\Big[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\Big] \le e^{\boldsymbol{\lambda}^\intercal \mathbf{C}\boldsymbol{\lambda}/2}. \tag{1}$$

Degenne and Perchet (2016) devised an algorithm with a regret bound depending on the components of another matrix $\boldsymbol{\Gamma} \succeq 0$, satisfying $\boldsymbol{\Gamma} \succeq_+ \mathbf{C}$ (i.e., $\boldsymbol{\lambda}^\intercal(\boldsymbol{\Gamma} - \mathbf{C})\boldsymbol{\lambda} \ge 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n_+$) and $\Gamma_{ij} \ge 0$ for all i, j. The major downside is that this algorithm *requires the knowledge* of $\boldsymbol{\Gamma}$. More precisely, their upper bound is of order

$$\frac{\log T}{\Delta} \sum_{i\in[n]} \Gamma_{ii}\big((1-\gamma)\log^2(m) + \gamma m\big), \tag{2}$$

where $\gamma \triangleq \max_{A\in\mathcal{A}} \max_{(i,j)\in A^2, i\neq j} \Gamma_{ij}/\sqrt{\Gamma_{ii}\Gamma_{jj}}$ is the the maximal off-diagonal correlation coefficient, Δ is the minimal positive gap between expected total reward of two actions, and $m \triangleq \max\{|A|,\ A \in \mathcal{A}\}$. Interestingly, their regret upper bound highlights regimes interpolating between worst case correlation between outcomes (corresponding to $\gamma = 1$) and mutually independent outcomes (where $\gamma = 0$). In particular, the learning rate is much faster in the latter case. The main drawback however, is that their approach is not adaptive since the correlation structure of the arms *needs to be given to the agent* (through the matrix $\boldsymbol{\Gamma}$).

Our main objective is to alleviate this issue, and to strive to obtain fast rates for combinatorial semi-bandits, as Degenne and Perchet (2016) in the case where there is a favorable covariance structure, but *without knowing* it beforehand. Therefore, algorithms should be able to capture the covariance structure given by $\mathbf{\Gamma}$ from the data processed and adapt to it. We actually go further by asking whether the matrix $\mathbf{\Gamma}$ is the relevant parameter to characterize the difficulty of a problem. We argue that the covariance matrix $\mathbf{\Sigma}^* \triangleq \mathbb{E}\big[(\mathbf{X}-\boldsymbol{\mu}^*)(\mathbf{X}-\boldsymbol{\mu}^*)^\intercal\big]$ is more pertinent, as it allows to better differentiate complex problems from the easy ones. One can indeed already argue in favor of a $\mathbf{\Sigma}^*$ dependence rather than a $\mathbf{\Gamma}$ one, based on the relation $\mathbf{\Sigma}^* \preceq_+ \mathbf{\Gamma}$ (see Appendix A).

Results and limitations of the results of Degenne and Perchet (2016) Below, we list the main limitations of the approach of Degenne and Perchet (2016):

- (i) The matrix $\mathbf{\Gamma}$ needs to be known. This requires specific knowledge about the outcome structure, which is often not precise, as it is usually only known that outcomes are bounded, or at most that there exists some constant κ such that $\kappa^2 \geq C_{ii}$ for all $i \in [n]$. The latter is equivalent[1] to $\Gamma_{ij} = \kappa^2$ for all $i, j \in [n]$ and corresponds to the worst case correlation between outcomes ($\gamma = 1$) in the regret bound (2).
- (ii) The value γ can be 1, even when outcomes are only weakly correlated: For instance, if n is even, $\mathbf{\Gamma}$ can be a block-diagonal matrix with $n/2$ blocks of size 2×2 containing only ones. This scenario can actually occur in many examples; we provide two types below:
  - Arms are nodes on a given graph, with some small communities on which outcome tends to be constant (Cesa-Bianchi et al., 2013; Valko et al., 2014; Gentile et al., 2014; Valko, 2016).
  - Arms are market-basket-like items, with some highly correlated pairs of items (e.g., people buying from category "books" tend to also buy from category "CDs", Zhang and Feigenbaum, 2006; He et al., 2006).
- (iii) The value Γ_{ii} can be high, even for low-variance outcomes, while intuitively, low variance outcomes should be easy to work with. For example, if $\mathbf{X}$ is a binary 1-sparse random variable — as in some recommender systems, where a single item is desired by the user — then $X_i \sim \text{Bernoulli}(\mu_i^*)$ with $\sum_{i=1}^n \mu_i^* = 1$, and $\Gamma_{ii} \geq C_{ii} \geq (\mu_i^* - 1/2)/(\log(\mu_i^*) - \log(1-\mu_i^*))$ (and this is tight, see, e.g., Buldygin and Moskvichova, 2013). For μ_i^* of order $1/n$, Γ_{ii} is thus at least of order $1/(2\log n)$ for n large, whereas $\mathbb{V}(X_i)$ is of order $1/n$.

To sum up the arguments above, we claim that (1) knowing a good upper bound on the sub-Gaussian matrix $\mathbf{C} \preceq_+ \mathbf{\Gamma}$ is not realistic and (2) even this upper bound is not a good proxy for the complexity of the instance at hand.

Contributions In this paper, we address the three aforementioned criticisms (i), (ii), and (iii). As a consequence, we do not assume that a good upper bound $\mathbf{\Gamma}$ on the sub-Gaussian matrix $\mathbf{C}$ is known, but only that the agent knows that each marginal $\mathbb{P}_{X_i}$ is κ^2-sub-Gaussian. We compensate this relaxation by restricting the distribution family considered through a sub-exponential-type assumption involving the covariance matrix $\mathbf{\Sigma}^*$. We argue that this restriction is mild and satisfied by many outcome distributions, including bounded and Gaussian.

1. Indeed, $\mathbf{C} \preceq_+ \mathbf{\Gamma} \Rightarrow C_{ii} \leq \kappa^2$ for all $i \in [n] \Rightarrow C_{ij} \leq \sqrt{C_{ii}C_{jj}} \leq \kappa^2$ for all $i, j \in [n] \Rightarrow \sum_{i,j} C_{ij}|\lambda_i||\lambda_i| \leq \kappa^2\big(\sum_i |\lambda_i|\big)^2$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n \Rightarrow \mathbf{C} \preceq_+ \mathbf{\Gamma}$.

We characterize the difficulty of the problem with $\mathbf{\Sigma}^*$; specifically, we provide a new lower bound, with a dependence on $\mathbf{\Sigma}^*$, more precise than Degenne and Perchet (2016). We also design a new algorithm with matching asymptotic regret upper bound, improving over the state-of-the-art results. One of the key techniques is to build an online adapted estimation of the matrix $\mathbf{\Sigma}^*$.

Our main contribution is in the analysis of this approach, that is not based on the usual *Laplace's method*, which works in the sub-Gaussian framework, but does not handle well our sub-exponential-type assumption. Thus, our analysis is rather based on a *covering-argument* (Magureanu et al., 2014). An important part of our proof is based on the transformation of the axis-unaligned ellipsoidal confidence region associated to a given action $A \in \mathcal{A}$ into an axis-aligned region, using the following relation $\left(\Sigma^*_{ij}\right)_{ij\in A} \preceq_+ \operatorname{diag}\left(\sum_{j\in A} 0 \vee \Sigma^*_{ij}\right)_{i\in A}$. This allows us to conduct the same type of proof than for the independent outcome case (where confidence regions are always axis-aligned), but with a Bernstein-type analysis.[2]

We also consider an application of our approach to the family of sparse bounded outcomes: we provide a lower bound on the regret, with an algorithm having a matching asymptotic regret upper bound.

Prior work on stochastic semi-bandits We review algorithms for stochastic semi-bandits, coming with the analysis that depends on the family of probability distributions to which $\mathbb{P}_{\mathbf{X}}$ belongs. To begin, Kveton et al. (2015); Chen et al. (2016) studied the general family of distributions having sub-Gaussian or bounded marginals. Their algorithms are not adaptive to $\mathbb{P}_{\mathbf{X}}$ and regret bounds depend on parameters characterizing the family, that need to be known (such as the sub-Gaussian constant or a bound on $\|\mathbf{X}\|_\infty$). On the other hand, many algorithms are *only* adaptive to marginals of $\mathbb{P}_{\mathbf{X}}$, either with variance estimates (Perrault et al., 2019b; Merlis and Mannor, 2019), or using Kullback–Leibler divergence. These approaches are agnostic to possible correlation between marginals since the confidence region used in their algorithm are always a Cartesian product of confidence intervals (so they are always n-dimensional hypercubes). As a consequence, this translates into guarantees w.r.t. the worst-case correlations quantity possible. Notice that these algorithms are actually almost direct applications of corresponding classical multi-arm bandits algorithms to the semi-bandit setting. In particular, confidence regions considered are the same in both settings.

Another line of works restricts the probability distributions family of $\mathbb{P}_{\mathbf{X}}$, so that the dependence existing between arms is controlled. This conveniently induce better confidence regions valid for distributions in the family, and leads to the development of algorithms based on these regions, having sharper regret upper bounds. For instance, Combes et al. (2015) and Wang and Chen (2018); Perrault et al. (2020) assumed that $\mathbb{P}_{\mathbf{X}} = \otimes_{i\in[n]}\mathbb{P}_{X_i}$. Confidence regions resemble to axis-align ellipsoid in this specific case. They designed UCB (resp. Thompson sampling) based algorithms, leveraging on such tighter ellipsoidal confidence region. The key difference between the above case is that this time, marginals do characterize the problem, by assumption on the probability distributions family.

Remark that Degenne and Perchet (2016) provided a regret bound which adapts to the probability distribution family at hand through the matrix $\mathbf{\Gamma}$, although their algorithm is not fully adaptive. The confidence region used by their algorithm is also ellipsoidal, and depends on the matrix $\mathbf{\Gamma}$. This matrix gives the control on the correlations between arms. The confidence ellipsoid is not axis aligned unless $\mathbf{\Gamma}$ is diagonal. To the best of our knowledge, their work is the main competitor in terms of regret bound.

2. Remark that contrary to previous work on variance based confidence region, our method can't be easily generalized to Kullback–Leibler divergence based confidence region, since this would require control on higher moments of $\mathbf{X}$.

Sparse bandits Independently to combinatorial bandits, there exists a different setting actually dealing with correlated outcomes in online learning known as *sparse bandits* (Kwon et al., 2017; Kwon and Perchet, 2015; Bubeck et al., 2017; Abbasi-Yadkori et al., 2012; Carpentier and Munos, 2012; Gerchinovitz, 2013). The overall idea is to introduce by now a standard sparsity assumption (some parameter vector has only s out of its n components that are non zero) into sequential decision making. As usual, the objective is to replace the linear/polynomial dependence in the dimension n by a linear/polynomial dependence in s. Quite interestingly, the sparsity assumption has been studied in two different directions. The fist one assumes that the vector $\boldsymbol{\mu}^*$ is s-sparse, typically in (linear) stochastic bandits (Kwon et al., 2017; Abbasi-Yadkori et al., 2012; Carpentier and Munos, 2012; Gerchinovitz, 2013). The second one assumes that the realized vector $\mathbf{X}_t$ is s-sparse, usually in adversarial bandits (Kwon and Perchet, 2015; Bubeck et al., 2017).

Sparsity in realized outcomes naturally induces negative correlation; this is not necessarily true for sparsity in expectation. More generally both concepts are complementary, since $\boldsymbol{\mu}^*$ can be sparse with non-sparse realization (for instance, if all X_i are i.i.d., equal to ± 1 with probability $1/2$) and reciprocally (if $\mathbf{X}$ is a canonical unit vector at random, then its expectation has full support). Surprisingly, the sparse outcomes setting has not been investigated in stochastic bandits, even if it lies at the junction of several notions of correlations between outcomes.

2. Some technical background

Let $\mathbf{e}_i$ be the i^{th} canonical unit vector of $\mathbb{R}^n$. The incidence vector of any subset $A \subset [n]$ is $\mathbf{e}_A \triangleq \sum_{i \in A} \mathbf{e}_i$. The above definition allows us to represent a subset of $[n]$ as an element of $\{0,1\}^n$. We denote the Minkowski sum of two sets $Z, Z' \subset \mathbb{R}^n$ as $Z + Z' \triangleq \{z + z', z \in Z, z' \in Z'\}$, and $Z + z' \triangleq Z + \{z'\}$. In *stochastic combinatorial semi-bandits*, an agent selects an action $A_t \in \mathcal{A}$ at each round $t \in \mathbb{N}^*$, and receives a reward $\mathbf{e}_{A_t}^\intercal \mathbf{X}_t$, where $\mathbf{X}_t \in \mathbb{R}^n$ is an unknown random vector of outcomes. The successive vectors $(\mathbf{X}_t)_{t \geq 1}$ are i.i.d., sampled from $\mathbb{P}_{\mathbf{X}}$, with an unknown mean $\boldsymbol{\mu}^* \triangleq \mathbb{E}[\mathbf{X}] \in \mathbb{R}^n$. After selecting an action A_t in round t, the agent observes the outcome of each individual arm in A_t. Its goal is to minimize the regret, defined with $A^* \in \arg\max_{A \in \mathcal{A}} \mathbf{e}_A^\intercal \boldsymbol{\mu}^*$ as

$$\forall T \geq 1, \quad R_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T} (\mathbf{e}_{A^*} - \mathbf{e}_{A_t})^\intercal \mathbf{X}_t\right].$$

For any action $A \in \mathcal{A}$, we define its gap as the difference $\Delta(A) \triangleq (\mathbf{e}_{A^*} - \mathbf{e}_A)^\intercal \boldsymbol{\mu}^*$. We then rewrite the regret as $R_T = \mathbb{E}\left[\sum_{t=1}^{T} \Delta(A_t)\right]$. We start by stating the assumptions satisfied by $\mathbb{P}_{\mathbf{X}}$.

Assumption 1 (κ^2**-sub-Gaussian marginals)** *There is a constant $\kappa > 0$ (known to the agent) such that $\forall i \in [n], \forall \lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda(X_i - \mu_i^*)}\right] \leq e^{\kappa^2 \lambda^2 / 2}$.*

Assumption 1 is not difficult to satisfy, and does not require any precision on the correlations between outcomes. In particular, Assumption 1 includes Gaussian outcomes (with variance lower than κ^2) and bounded outcomes (with $\|\mathbf{X}\|_\infty \leq \kappa$). We also assume that $\mathbf{X}$ satisfies the following.

Assumption 2 ($\|\cdot\|_1$**-sub-exponential distribution)** *$\forall \boldsymbol{\lambda} \in \mathbb{R}^n$ such that $\|\boldsymbol{\lambda}\|_1 \leq 1/(2\kappa)$, we have $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal (\mathbf{X} - \boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^* \boldsymbol{\lambda}}$, where $\boldsymbol{\Sigma}^* \triangleq \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}^*)(\mathbf{X} - \boldsymbol{\mu}^*)^\intercal\right]$ is the covariance matrix of $\mathbf{X}$.*

Importantly, the agent does not know the covariance matrix $\mathbf{\Sigma}^*$. Remark that Assumption 2 trivially holds for $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$, where $\forall \boldsymbol{\lambda} \in \mathbb{R}^n$, $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] = e^{\boldsymbol{\lambda}^\intercal \mathbf{\Sigma}^* \boldsymbol{\lambda}/2}$. The following proposition, proved in Appendix B, states that it also holds for bounded outcomes.

Proposition 1 *If $\|\mathbf{X}\|_\infty \leq \kappa$, then both Assumption 1 and 2 hold.*

Notice, up to a re-normalization of the regret, we assume w.l.o.g. that $\kappa = 1$.

3. Lower bound

We start by proving in Theorem 1 a new gap-dependent lower bound on R_T, valid for any covariance matrix $\mathbf{\Sigma}^* \succeq 0$, for some $\mathbb{P}_\mathbf{X}$ satisfying Assumptions 1 and 2, some action space $\mathcal{A}$, and for any consistent algorithm (Lai and Robbins, 1985), for which the regret on any problem verifies $R_T = o(T^a)$ as $T \to \infty$, for all $a > 0$. This lower bound demonstrates the link between $\mathbf{\Sigma}^*$ and the difficulty of the problem. It also indicates, in anticipation, that we have to examine a subclass of action sets to hope to improve the upper bound we will provide in Theorem 2.

Theorem 1 *For any $n, m \in \mathbb{N}^*$ such that $n/m \geq 2$ is an integer, any $n \times n$ matrix $\mathbf{\Sigma}^* \succeq 0$, any $\Delta > 0$, and any consistent policy, there exists an instance with n arms — characterized by some action space $\mathcal{A}$, with $m = \max\{|A|,\ A \in \mathcal{A}\}$, some outcome distribution $\mathbb{P}_\mathbf{X}$ satisfying Assumptions 1 and 2 with all gaps equal to Δ and covariance matrix $\mathbf{\Sigma}^*$ — on which the regret satisfies*

$$\liminf_{T\to\infty} \frac{\Delta}{\log(T)} R_T \geq 2 \sum_{i\in[n],\ i\notin A^*} \max_{A\in\mathcal{A},\ i\in A} \sum_{j\in A} \Sigma^*_{ij}.$$

The proof is given in Appendix C and considers $\mathcal{A}$ containing n/m disjoint actions $A_1, \dots, A_{n/m}$ composed of m arms, with $A_k = \{(k-1)m+1, \dots, km\}$, and $\mathbf{X} \sim \mathcal{N}(-\Delta/m(\mathbb{I}\{i \notin A_1\})_i, \mathbf{\Sigma}^*)$. The idea is to make a reduction to some standard bandit problems with n/m arms, and to compute the number of rounds t needed to distinguish between A_k and A_1. Roughly speaking, t is at least equal to the inverse of the KL between outcome distributions of A_k and its centered version, and in the case of Gaussian distributions, we get $t \geq 2\mathbb{V}(\sum_{i\in A_k} X_i)/\Delta^2 = 2\mathbf{e}_{A_k}^\intercal \mathbf{\Sigma}^* \mathbf{e}_{A_k}/\Delta^2$. It is not surprising that the variance appears, since this can be seen as a measure of the uncertainty we have in our samples: The higher the variance, the harder the estimation, and therefore the higher the round t must be. Notice that Theorem 1 is a refinement of Theorem 1 from Degenne and Perchet (2016), in which they consider the same action space $\mathcal{A}$ but a specific choice for the matrix $\mathbf{\Sigma}^*$: it is a block-diagonal matrix with n/m blocks, where each block (corresponding to an action A) is equal to $\sigma^2((1-\gamma)\mathrm{diag}(\mathbf{e}_A) + \gamma \mathbf{e}_A \mathbf{e}_A^\intercal)$, i.e., they take the worst case correlation under the controls given by σ^2 and γ, and knowing that the problem given by $\mathcal{A}$ is agnostic to the correlations between the arms of two different blocks.

In the next section, we describe our algorithm ESCB-C (Algorithm 1) and provide an upper bound on its regret in Theorem 2, where the expression $\max_{A\in\mathcal{A}, i\in A} \sum_{j\in A} 0 \vee \Sigma^*_{ij}$ appears. Notice that this is very close to the expression given in Theorem 1. In fact, both expressions coincide when $\mathbf{\Sigma}^*$ has only non-negative entries.

4. Main algorithm and the guarantees

In this section, we present an algorithm for the setting introduced in Section 2. The method is stated as Algorithm 1. To find the action with the highest mean, the agent estimates the mean μ_i^* of every

arm i with their corresponding *empirical averages* defined as $\overline{\mu}_{i,t-1} \triangleq \sum_{u\in[t-1]} \frac{\mathbb{I}\{i\in A_u\}X_{i,u}}{N_{i,t-1}}$, for $t \geq 1$, where $N_{i,t-1} \triangleq \sum_{u\in[t-1]} \mathbb{I}\{i \in A_u\}$ is the number of time arm i have been drawn for the first $t-1$ rounds. As mentioned above, the agent also estimates the covariance $\Sigma^*_{ij} = \mathbb{E}[X_i X_j] - \mu^*_i \mu^*_j$ of each pair $(i,j) \in [n]^2$. This will be done with the following estimate

$$\begin{aligned}\overline{\Sigma}_{ij,t-1} &\triangleq \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}\big(X_{i,u} - \overline{\mu}_{i,t-1}\big)\big(X_{j,u} - \overline{\mu}_{j,t-1}\big)}{N_{ij,t-1}} \\ &= \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}\big(X_{i,u}X_{j,u} - \overline{\mu}_{i,t-1}X_{j,u} - \overline{\mu}_{j,t-1}X_{i,u}\big)}{N_{ij,t-1}} + \overline{\mu}_{i,t-1}\overline{\mu}_{j,t-1},\end{aligned}$$

where $N_{ij,t-1} \triangleq \sum_{u\in[t-1]} \mathbb{I}\{i,j \in A_u\}$ is the number of times where arm i and j have been drawn *together* for the first $t-1$ rounds. Notice that in order to efficiently update $\overline{\Sigma}_{ij,t-1}$, in addition to $\overline{\mu}_{i,t-1}$ and $\overline{\mu}_{i,t-1}$, we only have to maintain the three quantities,

$$\sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}X_{i,u}X_{j,u}}{N_{ij,t-1}}, \quad \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}X_{i,u}}{N_{ij,t-1}}, \quad \text{and} \quad \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}X_{j,u}}{N_{ij,t-1}}.$$

Using concentration inequalities, we get confidence intervals for the above estimates. We are then able to use an upper-confidence-bound strategy (Auer et al., 2002). More precisely, we first build the upper confidence bound on Σ^*_{ij} using the fact that $X_i \cdot X_j$ is a *sub-exponential* random variable, since both X_i and X_j are sub-Gaussian by virtue of Assumption 1. The result is stated in the following proposition with a proof in Appendix D.

Proposition 2 *With probability* $1 - 10t^{-2}$*, we have*

$$\big|\Sigma^*_{ij} - \overline{\Sigma}_{ij,t-1}\big| \leq g_{ij}(t) \triangleq 16\left(\frac{3\log(t)}{N_{ij,t-1}} \vee \sqrt{\frac{3\log(t)}{N_{ij,t-1}}}\right) + \sqrt{\frac{48\log^2(t)}{N_{ij,t-1}N_{i,t-1}}} + \sqrt{\frac{36\log^2(t)}{N_{ij,t-1}N_{j,t-1}}}.$$

In particular, defining the upper confidence bound $\Sigma_{ij,t} \triangleq \overline{\Sigma}_{ij,t-1} + g_{ij}(t)$*, it holds that* $0 \leq \Sigma_{ij,t} - \Sigma^*_{ij} \leq 2g_{ij}(t)$ *with probability* $1 - 10t^{-2}$.

To build estimates well concentrated around $\boldsymbol{\mu}^*$, we will use the matrix $\boldsymbol{\Sigma}_t$ defined above to design the following high probability confidence region for all $A \in \mathcal{A}$

$$\mathcal{C}_t(A) \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{\boldsymbol{\xi} \in \mathbb{R}^n, \sum_{i\in A} \frac{N_{i,t-1}\xi_i^2}{|A||\xi_i| + \sum_{j\in A} 0 \vee \Sigma_{ij,t}} \leq 8(\log t + \log\log t) + 4em\right\}. \tag{3}$$

The intuition behind this confidence region is similar to the one for empirical Bernstein confidence intervals, but the term $\sum_{j\in A} 0 \vee \Sigma_{ij,t}$ in the denominator replaces the usual empirical variance. To compare our confidence region with the one of Degenne and Perchet (2016), notice first that their algorithm uses the matrix $\boldsymbol{\Gamma}$ to build a confidence ellipsoid. They provide an analysis for this confidence ellipsoid using the *Laplace's method* and the matrix relation $\mathbf{C} \preceq_+ \boldsymbol{\Gamma}$. In contrast, our confidence region is based on the covariance matrix $\boldsymbol{\Sigma}^*$. Our analysis is also different, as we use a *covering-argument* analysis. This is because the covariance estimation and Assumption 2 are both hard to handle with Laplace's method, that is more appropriate for sub-Gaussian random

Algorithm 1 ESCB-C *(Efficient Sampling for Combinatorial Bandits with Covariance estimate)*

Initialization:
Play $A_1 = [n]$, or at least a sequence $A_1, A_2, \dots$, (no more than $n(n-1)/2$) such that for any $i, j \in [n]$, one of these A_t's contains $\{i, j\}$. We thus have $N_{ij,t-1} \geq 1$ for all $i, j \in [n]$.
For all subsequent rounds t:
Solve the following bilinear program to get A_t, with $\mathcal{C}_t(A)$ defined by (3), and play A_t,

$$(A_t, \boldsymbol{\mu}_t) \in \underset{A \in \mathcal{A},\ \boldsymbol{\mu} \in \mathcal{C}_t(A)}{\arg\max} \mathbf{e}_A^\mathsf{T} \boldsymbol{\mu}.$$

variables. Indeed, all calculations can be explicit and it is easy to construct a *conjugate prior*. This is *not the case* for sub-exponential random variables. Covering arguments are much more easier to use together with a diagonal matrix, so axis-aligned confidence region are desirable. We use an *axis-realignment technique* based on the matrix relation $\left(\Sigma^*_{ij}\right)_{ij \in A} \preceq_+ \operatorname{diag}\left(\sum_{j \in A} 0 \vee \Sigma^*_{ij}\right)_{i \in A}$. The upside is to avoid dealing with off-diagonal terms by transforming them into diagonal ones. From all these previous observations, we can say that the confidence ellipsoid of Degenne and Perchet (2016) is tighter as it does not require any axis realignment; however, not only the matrix $\boldsymbol{\Gamma}$ is generally looser than $\boldsymbol{\Sigma}^*$ but also axis realignment does not alter the analysis, so that our new approach outperforms theirs in terms of asymptotic regret upper bound.

As common in bandits, the major challenge in the analysis is to prove that with high probability, $\boldsymbol{\mu}^* \in \mathcal{C}_t(A)$ for any action $A \in \mathcal{A}$. The covering argument together with the conversion from an axis-unaligned confidence region into an axis-aligned confidence region allows us to achieve this result (see Lemma 3). Therefore, an optimistic estimate $\boldsymbol{\mu}_t$ of the true mean $\boldsymbol{\mu}^*$ can be found using an upper-confidence-bound approach: if $A_t, \boldsymbol{\mu}_t$ are defined as in Algorithm 1, then, since $\boldsymbol{\mu}^* \in \mathcal{C}_t(A^*)$, we have

$$\mathbf{e}_{A_t}^\mathsf{T} \boldsymbol{\mu}_t \geq \mathbf{e}_{A^*}^\mathsf{T} \boldsymbol{\mu}^*.$$

The regret bound for ESCB-C is stated in Theorem 2 with proof in Appendix E.

Theorem 2 *Assume that the outcome distribution $\mathbb{P}_\mathbf{X}$ satisfies Assumptions 1 and 2, and define $\Delta \triangleq \min_{A \in \mathcal{A},\ \Delta(A) > 0} \Delta(A)$, $\Delta_{\max} \triangleq \max_{A \in \mathcal{A},\ \Delta(A) > 0} \Delta(A)$. If Δ is small enough, i.e., there exists a universal constant c such that*

$$\Delta \vee \left(\Delta + \Delta \log\left(\frac{\Delta_{\max}}{\Delta}\right)\right)^{3/2} \leq c \left(\frac{\log(m+1) \sum_{i \in [n]} \max_{A \in \mathcal{A}, i \in A} \sum_{j \in A} 0 \vee \Sigma^*_{ij}}{n^2}\right)^{3/2},$$

then the regret of Algorithm 1 is upper bounded as

$$\limsup_{T \to \infty} \frac{\Delta}{\log T} R_T \leq c' \log^2(m+1) \sum_{i \in [n]} \max_{A \in \mathcal{A}, i \in A} \sum_{j \in A} 0 \vee \Sigma^*_{ij},$$

where c' is a universal constant.

Notice that the bound in Theorem 2 is tight, up to a poly-logarithmic factor in m, with respect to the lower bound in Theorem 1, in the case where $\boldsymbol{\Sigma}^*$ has non-negative entries. Moreover, we focus on the asymptotic behavior of the regret (w.r.t. T) when Δ is small, i.e., when the problem becomes very difficult. While the quantity $c \log^2(m+1) \sum_{i \in [n]} \max_{A \in \mathcal{A}, i \in A} \sum_{j \in A} 0 \vee \Sigma^*_{ij} \log(T)/\Delta$ presented in Theorem 2 highlights the main dependence on both Δ and T, we prove a more precise

non-asymptotic upper bound in Appendix E, (9), which holds for all $\Delta > 0$. Indeed, as for UCB-V, the errors from estimating $\mathbf{\Sigma}$ generate an extra term in the upper bound. However, since these errors are multiplied with estimation errors on the means, their impact is of second order. In particular, for Δ small enough, this extra term becomes negligible compared to the main term. Therefore, the term from covariance estimation errors is *not* present in Theorem 2, but appears when Δ is far from 0. Finally, remark that when the covariance $\mathbf{\Sigma}^*$ is known, then one can consider the confidence region where $\Sigma_{ij,t}$ is replaced by Σ^*_{ij}. This avoids covariance estimation errors, and gives the upper bound of Theorem 2 when $\Delta + \Delta \log(\Delta_{\max}/\Delta)$ is smaller than $\sum_{i\in[n]} \max_{A\in\mathcal{A}, i\in A} \sum_{j\in A} 0 \vee \frac{\Sigma^*_{ij}}{n\cdot m}$.

Remark 1 *Considering the intersection of the region from Algorithm 1 with the one of* CUCB-V, *we can replace* $\log^2(m+1)\sum_{j\in A} 0 \vee \Sigma^*_{ij}$ *by* $m\Sigma^*_{ii} \wedge \left(\log^2(m+1)\sum_{j\in A} 0 \vee \Sigma^*_{ij}\right)$ *in Theorem 2.*

5. Application to sparse outcomes

In this section, we shall consider an additional structural assumption on the vector $\mathbf{X}$, namely that it is s-sparse in the sense that

$$\|\mathbf{X}\|_0 \leq s,$$

i.e., the number of nonzero components of $\mathbf{X}$ is smaller than s, where s is a fixed known parameter.[3] Importantly, the set of components which are nonzero is *not fixed nor known*, and may change over time. It should be noted, however, that there is a significant difference between the stochastic and the adversarial cases: In the later, the set of components which are nonzero change arbitrarily over time, whereas in the former, this set is sampled i.i.d. Notice, this sparse stochastic setting is different than the usual stochastic sparse bandit, where $\boldsymbol{\mu}^*$ is assumed to be sparse; see e.g., Kwon et al. (2017) for the classical MAB setting, and Abbasi-Yadkori et al. (2012); Carpentier and Munos (2012) for the linear bandit setting. For simplicity, we further assume that $\|\mathbf{X}\|_\infty \leq 1$. As we already saw in Proposition 1, this implies Assumption 1 and 2. The difficulty of this setting is that both the approach of Degenne and Perchet (2016) and standard methods such as CUCB-V would not reach the lower bound for the regime $s \leq m$, as we will see. The reason is that a correlation exists between the components, because of sparsity, and must be taken into account.

Why sparsity in semi-bandits? Sparsity is nowadays a very standard assumption in learning theory (that potentially does not need any further motivations). There are many examples of online learning scenarios naturally involving some sparse structure. For instance, in the celebrated click-through-rate optimization, it is safe to assume that users would only click on a few of the different ads that can be displayed (those that can catch their eyes for any reason, say). Similarly, in recommender systems, it is safe to assume that a user will browse/buy items from a specific category and not the other (for instance, a segment of the population in e-shops only buy bottles of wines and others only video-games or clothes).

Other examples involve settings where outcomes are usually zero except on very rare occasions: In the online routing, the packets are sent in a network and are lost if a server of that network has a failure. Because of failsafe procedures, failures are *desynchronized* and typically only one (or at most a few) of them can happen simultaneously. In all of these examples, the decision maker has some combinatorial problem to solve: select an admissible path, select a *diverse* bundle of object/ads to display, etc., and only a few of the base items will generate non-zero outcome.

3. For example, the Dirichlet-multinomial distribution with s trials is s-sparse.

5.1. Lower bound

To start our study of sparse outcomes, we state a new lower bound in Theorem 3, that is valid for the setting described above. This lower bound is built on the same ideas as Theorem 1, with a notable variation: when reducing to a MAB problem, we do not obtain the necessary conditions for the application of Lai and Robbins (1985), because of the linear dependence between the μ_i^*'s. Thus, we use instead the the lower bound from Graves and Lai (1997). More precisely, we consider the same action space $\mathcal{A}$, and incorporate the sparsity assumption as an extra constraint for defining a worst case distribution.

Theorem 3 *For any $n, m, s \in \mathbb{N}^*$ such that n/m, n/s, $1 \vee (s/m)$ are integers, $n/m, n/s \geq 2$, any $\Delta \in (0, \frac{ms}{2(n-m)}]$ and any consistent policy, there is a problem with n arms — characterized by some action space $\mathcal{A}$ with $m = \max\{|A|,\ A \in \mathcal{A}\}$ and some vector of outcomes $\mathbf{X}$ with all gaps equal to Δ satisfying $\|\mathbf{X}\|_\infty \leq 1,\ \|\mathbf{X}\|_0 \leq s$ — on which the regret satisfies*

$$\liminf_{T\to\infty} \frac{\Delta}{\log(T)} R_T \geq \frac{s(s \wedge m)(1 - 2m/n)}{4}.$$

The proof is given in Appendix G. To give an idea, contrary to Theorem 1, we have more freedom in the covariance, and $\mathbf{X}$ can be chosen to maximize $\mathbb{V}(\sum_{i\in A} X_i)$ for each action A, up to the constraints $\|\mathbf{X}\|_\infty \leq 1,\ \|\mathbf{X}\|_0 = s$. The maximal value of $\sum_{i\in A} X_i$ is thus $(s \wedge m)$. Now consider for simplicity the softer constraint $\mathbb{E}\|\mathbf{X}\|_0 = s$. If $\mathbf{X}$ is chosen so that $\sum_{i\in A} X_i/(s\wedge m)$ is Bernoulli of parameter p, then the optimal p is equal to $(s \vee m)/n$. The variance is about $p(s \wedge m)^2 = ms(s \wedge m)/n$. Multiplying this by n/m (the number of actions) and dividing by the gap Δ gives the order of the lower bound.

5.2. Our approach for sparse semi-bandits

In this subsection, we adapt our techniques to the sparse semi-bandit setting. Since $\|\mathbf{X}\|_\infty \leq 1$, the ℓ_0-inequality $\|\mathbf{X}\|_0 \leq s$ immediately implies the ℓ_1-inequality $\|\mathbf{X}\|_1 \leq s$. As we will actually only use sparsity through the latter inequality, we can relax our assumption on the model into $\|\mathbf{X}\|_1 \leq s$, for more generality. Let $\nu_i^* \triangleq \mathbb{E}[|X_i|]$, and $\overline{\nu}_{i,t-1}$ the corresponding empirical average estimate: $\overline{\nu}_{i,t-1} \triangleq \frac{\sum_{u\in[t-1]} \mathbb{I}\{i\in A_u\}|X_{i,u}|}{N_{i,t-1}}$. Our approach is based on replacing $\sum_{j\in A} 0 \vee \Sigma_{ij}^*$ by $\nu_i^*(s \wedge m)$ (see Lemma 1, proved in Appendix H). Using this, it is possible to estimate ν_i^* instead of each Σ_{ij}^*.

Lemma 1 $\sum_{j\in A} 0 \vee \Sigma_{ij}^* \leq 2\nu_i^*(s \wedge m)$.

We can therefore use the same algorithm (Algorithm 1), but with a confidence region $\mathcal{C}_t$ independent of A, since summing over A or $[n]$ on the main sum doesn't change the algorithm and the second sum $\sum_{j\in A} 0 \vee \Sigma_{ij,t}$ is replaced by an estimates of the upper bound given in Lemma 1.

$$\mathcal{C}_t \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{ \boldsymbol{\xi} \in \mathbb{R}^n, \sum_{i\in[n]} \frac{N_{i,t-1}\xi_i^2}{m|\xi_i| + 2(s\wedge m)\nu_{i,t}} \leq 8(\log(t) + \log(\log(t))) + 4em \right\}, \tag{4}$$

where the upper bound estimate $\nu_{i,t} \triangleq \overline{\nu}_{i,t-1} + \sqrt{\frac{1.5\log(t)}{N_{i,t-1}}}$ of ν_i^* is a simple consequence of Hoeffding's inequality, using that $|X_{i,u}|$ is $1/4$-sub-Gaussian. Our algorithm is stated in Algorithm 2. As a byproduct of Theorem 2, we provide an upper bound for the regret in the sparse semi-bandit setting in Corollary 1 (see Appendix F, (10), for a more precise bound). Again, notice we are reaching the lower bound of Theorem 3, using the relation $\sum_i \nu_i^* = \mathbb{E}\|\mathbf{X}\|_1 \leq s$.

Algorithm 2 ESCB-C modified for the case of $\|\cdot\|_1$-constrained outcomes

Initialization:
Play $A_1 = [n]$, or at least a sequence $A_1, A_2, \dots,$ (no more than n) such that all arm have been sampled once. We thus have $N_{i,t-1} \geq 1$ for every arm $i \in [n]$.
For all subsequent rounds t:
Solve the following bilinear program to get A_t, with $\mathcal{C}_t$ define by (4), and play A_t.

$$(A_t, \boldsymbol{\mu}_t) \in \operatorname*{arg\,max}_{A \in \mathcal{A},\ \boldsymbol{\mu} \in \mathcal{C}_t} \mathbf{e}_A^\mathsf{T} \boldsymbol{\mu}.$$

Corollary 1 *Assume that the outcome distribution $\mathbb{P}_\mathbf{X}$ satisfies $\|\mathbf{X}\|_\infty \leq 1$ and $\|\mathbf{X}\|_1 \leq s$, and that*

$$(\Delta(s \wedge m))^{2/3} \vee (m\Delta + m\Delta \log(\Delta_{\max}/\Delta)) \leq c \log(m+1) \sum_{i \in [n]} \frac{\nu_i^*(s \wedge m)}{n},$$

for some universal constant c. Then the regret of Algorithm 2 is upper bounded as

$$\limsup_{T \to \infty} \frac{\Delta}{\log(T)} R_T \leq c' \log^2(m+1) \sum_{i \in [n]} \nu_i^*(s \wedge m) \leq c' \log^2(m+1)(s \wedge m)s,$$

where c' is a universal constant.

Remark 2 *It should be noticed that semi-bandits algorithms as* CUCB-V *or* CUCB-KL *(that are variant of the classical* CUCB *(Kveton et al., 2015), where the confidence region is a Cartesian product of confidence intervals, with Bernstein and kl-base confidence intervals respectively) also reach the lower bound of Theorem 3 for the regime $s \geq m$, since $\mathbb{V}(X_i) \leq 2\nu_i^*$ (thanks to Lemma 1). However, in the regime where $s \leq m$, these algorithms are not able to reach it, while* ESCB-C *is. In Appendix I, we describe the two algorithms* CUCB-V *and* CUCB-KL*, and comment further on the tightness difference between confidence regions.*

6. Implementation details

We now discuss the computational efficiency of our approaches. First, Algorithm 1 (and both those of Combes et al. (2015) and Degenne and Perchet (2016)) is not efficient for arbitrary combinatorial space $\mathcal{A}$. However, the evaluation of $F : A \mapsto \max_{\boldsymbol{\mu} \in \mathcal{C}_t(A)} \mathbf{e}_A^\mathsf{T} \boldsymbol{\mu}$, can be done efficiently as it is an LP over a convex set. In practice, when $\mathcal{A}$ allows it, GREEDY[4] (Nemhauser et al., 1978) can be used to maximize F. In general, it is unknown if this alters the regret rate. On the one hand, it does not when $\mathcal{A}$ is given by a matroid, and $\mathcal{C}_t$ is as in Algorithm 2. This is because F is *submodular* and the following approximation guaranty holds for the output A_t of GREEDY (Perrault et al., 2019a): $2\big(F(A_t) - \mathbf{e}_{A_t}^\mathsf{T} \overline{\boldsymbol{\mu}}_{t-1}\big) + \mathbf{e}_{A_t}^\mathsf{T} \overline{\boldsymbol{\mu}}_{t-1} \geq F(A^*)$, where the l.h.s. is simply F where $\mathcal{C}_t$ is scaled by a factor 2 from its center $\overline{\boldsymbol{\mu}}_{t-1}$. On the other hand, when $\mathcal{C}_t(A)$ is as in Algorithm 1, a concave extension of $A \mapsto F(A)$ can be considered, and can thus be maximized efficiently. Notice, when considering the intersection of the two confidence regions as in Remark 1, this implementation is still tractable since the minimum of two concave functions is still concave. Since the obtained solution might not be fractional, we use a randomized rounding to obtain a feasible set $A_t \in \mathcal{A} = \{0,1\}^n$. We provide in Appendix J further details and prove that this method scales the regret by a factor $1 + \log\left(\frac{m \log(T)}{\Delta^2}\right)$, an acceptable price for efficiency.

4. Starting from $A = \emptyset$, we sequentially add (when possible) the best possible i to the current A if $F(A \cup \{i\}) > F(A)$.

7. Experiments

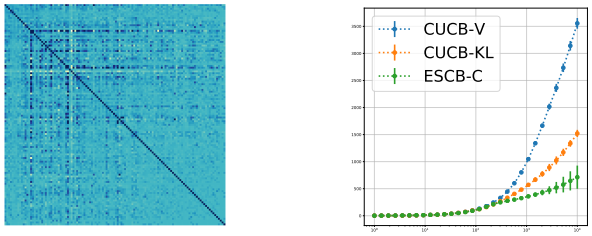

Figure 1: **Left**: Correlation matrix of the dataset, **right**: Cumulative regret, averaged over 36 independent simulations

We consider the following dynamic assortment problem. An agent has n products to sale, with fixed known prices. At each round, a customer arrives, with some unknown random valuation vector over products. Then, the agent offers any subset of products, by paying a fixed known cost for each offered product (e.g. transportation and display cost), and the customer buys an offered product if and only if its valuation is greater than its price. The agent is interested in maximizing the total profit (revenue minus cost) from sales over T rounds. We use the $n = 120$ products from the Kaggle Market Basket Optimization (2013) dataset containing 7500 grocery store transactions. At each round, valuations are determined by sampling a random transaction from this dataset. The choice of such data is motivated by correlations that exist between arms, as illustrated in Figure 1 – left, representing the correlation matrix. We ran 36 independent simulations with $T = 10^4$, and with a common product price and cost respectively equal to 1.5 and 0.1. We compared CUCB-V and CUCB-KL (see Appendix I) with the Lovász extension implementation of ESCB-C (see Appendix J) and results are plotted in log-scale (Figure 1 – right); error bars represent the sample standard deviation over simulations. There is less volatility in the regret of CUCB-V and CUCB-KL; this is due to the fact that their confidence regions overestimate the risk, and the "bad" event where the regret deviates is almost negligible. Nevertheless, we clearly observe that ESCB-C outperforms the two other approaches in terms of the average regret. Finally, let us point out that we did not empirically compare to the OLS-UCB algorithm of Degenne and Perchet (2016) since it is inefficient to implement (the combinatorial problem to be solved within each round is NP-Hard in general (Atamtürk and Gómez, 2017). We noticed that for the choice of sub-Gaussian matrix where all the correlation coefficients equals 1, OLS-UCB (if it could be implemented) would return a solution very close to CUCB-V.

8. Discussion

We improved the analysis of combinatorial semi-bandits in multiple ways. First, we brought new perspectives by considering a fairly large family of sub-exponential probability distributions, that crucially do not depend on parameters difficult to obtain in real situations. We have built an algorithm for this family, based on the estimation of the covariance matrix. We have therefore already significantly improved existing approaches by adapting not only to the variance of the arms, but also to the correlation between them. A tight analysis of our proposed method gives a new state-of-the-art upper bound on the regret. Our new bound is also more intuitive, and is more relevant to reflect the complexity of the instance at hand (through correlations between arms). Finally, we applied our approach to a setting not yet studied before, that assumes sparsity of the outcome vector. We gave a lower bound, as well as a matching algorithm that leverages the sparsity assumption.

Acknowledgments

The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, French National Research Agency project BOLD (ANR19-CE23-0026-04). Furthermore, it was also supported in part by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH, in a joint call with Gaspard Monge Program for optimization, operations research and their interactions with data sciences.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1–9, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/abbasi-yadkori12.html.

Alper Atamtürk and Andrés Gómez. Maximizing a class of utility functions over the vertices of a polytope. *Operations Research*, 65(2):433–445, 2017.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. URL https://homes.di.unimi.it/{~}cesabian/Pubblicazioni/ml-02.pdf.

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. *CoRR*, abs/1711.01037, 2017. URL http://arxiv.org/abs/1711.01037.

V. V. Buldygin and K. K. Moskvichova. The sub-Gaussian norm of a binary random variable. *Theory of Probability and Mathematical Statistics*, 86:33–49, 2013. ISSN 0094-9000. doi: 10.1090/s0094-9000-2013-00887-4.

Apostolos N. Burnetas and Michaël N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Alexandra Carpentier and Remi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 190–198, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/carpentier12.html.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. In *Journal of Computer and System Sciences*, volume 78, pages 1404–1422, 2012. URL http://cesa-bianchi.di.unimi.it/Pubblicazioni/comband.pdf.

Nicolò Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. A gang of bandits. In *Neural Information Processing Systems*, 2013. URL https://papers.nips.cc/paper/5006-a-gang-of-bandits.pdf.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17, 2016.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Others. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. dec 2016. URL https://arxiv.org/abs/1612.01859.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, 2011. URL https://arxiv.org/pdf/1102.2490.pdf.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, 2014. URL http://proceedings.mlr.press/v32/gentile14.pdf.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

A György, T Linder, G Lugosi, and Ottucsák. The On-Line Shortest Path Problem Under Partial Monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007. ISSN 1532-4435.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Mining top-k strongly correlated item pairs without minimum correlation threshold. *KES Journal*, 10:105–112, 03 2006. doi: 10.3233/KES-2006-10202.

Jean Honorio and Tommi Jaakkola. Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 384–392, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL http://proceedings.mlr.press/v33/honorio14.html.

Kaggle Market Basket Optimization. Kaggle, 2013. URL https://www.kaggle.com/roshansharma/market-basket-optimization.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. *Algorithmic Learning Theory*, 2012. URL https://arxiv.org/pdf/1205.4217.pdf.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, 2015. URL http://proceedings.mlr.press/v38/kveton15.pdf.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *CoRR*, abs/1511.08405, 2015. URL http://arxiv.org/abs/1511.08405.

Joon Kwon, Vianney Perchet, and Claire Vernade. Sparse stochastic bandits. *CoRR*, abs/1706.01383, 2017. URL http://arxiv.org/abs/1706.01383.

Tze L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. URL https://ac.els-cdn.com/0196885885900028/1-s2.0-0196885885900028-main.pdf?{_}tid=6ded14a5-1fe6-4c09-a1e3-9738a40b46d4{&}acdnat=1539373065{_}3220aa4053ab6e1f5db385fd4ef37e61.

László Lovász. Submodular functions and convexity. *Mathematical programming the state of the art*, pages 235–257, 1983. URL http://www.cs.elte.hu/{%}7B{~}{%}7Dlovasz/scans/submodular.pdf.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz Bandits: Regret Lower Bounds and Optimal Algorithms. may 2014. URL https://arxiv.org/abs/1405.4758.

Nadav Merlis and Shie Mannor. Batch-Size Independent Regret Bounds for the Combinatorial Multi-Armed Bandit Problem. may 2019. URL https://arxiv.org/abs/1905.03125.

Subhojyoti Mukherjee, K. P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-UCBV: An Almost Optimal Algorithm using Variance Estimates. nov 2017. URL https://arxiv.org/abs/1711.03591.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 14(1):265–294, 1978. ISSN 00255610. doi: 10.1007/BF01588971.

Pierre Perrault, Vianney Perchet, and Michal Valko. Exploiting structure of uncertainty for efficient matroid semi-bandits. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5123–5132, Long Beach, California, USA, 09–15 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v97/perrault19a.html.

Pierre Perrault, Vianney Perchet, and Michal Valko. Finding the bandit in a graph: Sequential search-and-stop. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1668–1677. PMLR, 16–18 Apr 2019b. URL http://proceedings.mlr.press/v89/perrault19a.html.

Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:2006.06613*, 2020.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

M. Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson. Stochastic Online Shortest Path Routing: The Value of Feedback. sep 2013. URL https://arxiv.org/abs/1309.7367.

Michal Valko. *Bandits on graphs and structures*. habilitation, École normale supérieure de Cachan, 2016.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, 2014.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Neural Information Processing Systems*, mar 2017a. URL http://arxiv.org/abs/1703.01610.

Qinshi Wang and Wei Chen. Tighter Regret Bounds for Influence Maximization and Other Combinatorial Semi-Bandits with Probabilistically Triggered Arms. *arXiv preprint arXiv:1703.01610*, 2017b.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. *CoRR*, abs/1803.04623, 2018. URL http://arxiv.org/abs/1803.04623.

Yingfei Wang, Hua Ouyang, Chu Wang, Jianhui Chen, Tsvetan Asamov, and Yi Chang. Thompson Sampling for Contextual Combinatorial Bandits. In *WOODSTOK '97*, 1997. doi: 10.475/123. URL http://asamov.com/download/ts{_}combinatorial.pdf.

Jian Zhang and Joan Feigenbaum. Finding highly correlated pairs efficiently with powerful pruning. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 152–161, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: 10.1145/1183614.1183640. URL http://doi.acm.org/10.1145/1183614.1183640.

Appendix A. The sub-Gaussian matrix is an upper bound on the covariance matrix

The fact that $\boldsymbol{\Sigma}^* \preceq \mathbf{C}$ is well known and can be proved as follows: Fix $\mathbf{x} \in \mathbb{R}^n$. For any $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda \mathbf{x}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq e^{\lambda^2 \mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x}/2}$. The second order Taylor expansion in λ gives

$$\frac{\lambda^2}{2}\mathbb{E}\left[(\mathbf{x}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^2\right] + o(\lambda^2) \leq \frac{\lambda^2}{2}\mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x} + o(\lambda^2).$$

Dividing the inequality by λ^2, and letting $\lambda \to 0$ yields $\mathbb{E}\left[(\mathbf{x}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^2\right] \leq \mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x}$, i.e., $\boldsymbol{\Sigma}^* \preceq \mathbf{C}$.

Appendix B. Proof of Proposition 1

Proof Assumption 1 is a direct consequence of Hoeffding's Lemma. For Assumption 2, we have $\|\mathbf{X}-\boldsymbol{\mu}^*\|_\infty \leq 2\kappa$. For $\|\boldsymbol{\lambda}\|_1 \leq 1/(2\kappa)$, we have:

$$\begin{aligned}
\log \mathbb{E}\left[e^{\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*)}\right] &= \log\left(1 + \sum_{k\geq 2}\mathbb{E}\left[\frac{(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^k}{k!}\right]\right) \\
&\leq \sum_{k\geq 2}\mathbb{E}\left[\frac{(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^k}{k!}\right] && \log(x) \leq x-1 \ \forall x>0, \\
&= \sum_{k\geq 2}\mathbb{E}\left[\frac{(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^{k-2}(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^2}{k!}\right] \\
&\leq \sum_{k\geq 2}\mathbb{E}\left[\frac{(\|\boldsymbol{\lambda}\|_1\|\mathbf{X}-\boldsymbol{\mu}^*\|_\infty)^{k-2}(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^2}{k!}\right] \\
&\leq \sum_{k\geq 2}\frac{\mathbb{E}\left[(\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*))^2\right]}{k!} \\
&= (e-2)\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\Sigma}^*\boldsymbol{\lambda} \leq \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\Sigma}^*\boldsymbol{\lambda}.
\end{aligned}$$

■

Appendix C. Proof of Theorem 1

Proof Consider $\mathcal{A}$ containing n/m disjoint actions $A_1, \ldots, A_{n/m}$ composed of m arms, with $A_k = \{(k-1)m+1, \ldots, km\}$, and $\mathbf{X} \sim \mathcal{N}(-\Delta/m(\mathbb{I}\{i \notin A_1\})_i, \boldsymbol{\Sigma}^*)$. This problem reduces to a standard bandit problem with n/m arms. We use a result from Burnetas and Katehakis (1996), a generalization of Lai and Robbins (1985), that states that

$$\liminf_{T\to\infty}\frac{R_T}{\log(T)} \geq \sum_{k=2}^{n/m}\frac{\Delta}{\inf_{Y,\ \mathbb{E}[Y]=0}\mathrm{KL}\left(\mathbb{P}_{\sum_{i\in A_k}X_i}\|\mathbb{P}_Y\right)}.$$

As we can write

$$\inf_{Y,\ \mathbb{E}[Y]=0} \mathrm{KL}\Big(\mathbb{P}_{\sum_{i\in A_k} X_i} \| \mathbb{P}_Y\Big) \le \mathrm{KL}\big(\mathcal{N}\big(-\Delta, \mathbf{e}_{A_k}^\mathsf{T} \boldsymbol{\Sigma}^* \mathbf{e}_{A_k}\big) \| \mathcal{N}\big(0, \mathbf{e}_{A_k}^\mathsf{T} \boldsymbol{\Sigma}^* \mathbf{e}_{A_k}\big)\big)$$
$$= \frac{\Delta^2/2}{\mathbf{e}_{A_k}^\mathsf{T} \boldsymbol{\Sigma}^* \mathbf{e}_{A_k}},$$

it holds that

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \ge 2 \sum_{k=2}^{n/m} \frac{\mathbf{e}_{A_k}^\mathsf{T} \boldsymbol{\Sigma}^* \mathbf{e}_{A_k}}{\Delta} = 2 \sum_{i\in[n],\ i\notin A_1} \max_{A\in\mathcal{A},\ i\in A} \sum_{j\in A} \frac{\Sigma^*_{ij}}{\Delta},$$

where we used the fact that $\{A \in \mathcal{A},\ i \in A\}$ is a singleton. ■

Appendix D. Proof of Proposition 2

Proof We define $\widetilde{\Sigma}_{ij,t-1} \triangleq \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j\in A_u\}\big(X_{i,u}-\mu_i^*\big)\big(X_{j,u}-\mu_j^*\big)}{N_{ij,t-1}}$ and for $k \in \{i,j\}$, $\widetilde{\mu}_{k,t-1} \triangleq \frac{1}{N_{ij,t-1}} \sum_{u\in[t-1]} \mathbb{I}\{i,j \in A_u\} X_{k,u}$. Notice that the following relation holds

$$\overline{\Sigma}_{ij,t-1} = \widetilde{\Sigma}_{ij,t-1} + \big(\mu_i^* - \overline{\mu}_{i,t-1}\big)\big(\widetilde{\mu}_{j,t-1} - \overline{\mu}_{j,t-1}\big) + \big(\mu_j^* - \overline{\mu}_{j,t-1}\big)\big(\widetilde{\mu}_{i,t-1} - \mu_i^*\big).$$

We now state Lemma 2 giving sub-exponential parameters for a product of sub-Gaussian random variables. A proof comes from Honorio and Jaakkola (2014).

Lemma 2 *Y,Z are 1-sub-Gaussian random variables* $\Rightarrow \forall|\lambda| \le 1/8,\ \mathbb{E}\big[e^{\lambda(YZ-\mathbb{E}[YZ])}\big] \le e^{64\lambda^2}$.

We apply Lemma 2 with a Chernoff argument and an union bound (to avoid the randomness of counters) in order to get the following Bernstein inequality

$$\mathbb{P}\left[\left|\Sigma^*_{ij} - \widetilde{\Sigma}_{ij,t-1}\right| \ge 16\left(\frac{3\log(t)}{N_{ij,t-1}} \vee \sqrt{\frac{3\log(t)}{N_{ij,t-1}}}\right)\right] \le 2t^{-2}.$$

In the same way, Hoeffding's inequality gives directly that with probability $1 - 8t^{-2}$, we have simultaneously

$$\begin{cases} \big|\mu_i^* - \overline{\mu}_{i,t-1}\big| & \le \sqrt{\frac{6\log(t)}{N_{i,t-1}}} \\ \big|\widetilde{\mu}_{j,t-1} - \overline{\mu}_{j,t-1}\big| & \le \sqrt{\frac{8\log(t)}{N_{ij,t-1}}} \\ \big|\mu_j^* - \overline{\mu}_{j,t-1}\big| & \le \sqrt{\frac{6\log(t)}{N_{j,t-1}}} \\ |\widetilde{\mu}_{i,t-1} - \mu_i^*| & \le \sqrt{\frac{6\log(t)}{N_{ij,t-1}}}, \end{cases}$$

which is enough to conclude the proof. Notice that for the second inequality above, we take the union bound for two counters. When they are not random, $N_{ij,t-1}\big(\widetilde{\mu}_{j,t-1} - \overline{\mu}_{j,t-1}\big)$, that is equal to

$$\sum_{u\in[t-1]} \mathbb{I}\{i,j\in A_u\} X_{j,u}(1 - N_{ij,t-1}/N_{j,t-1}) - \sum_{u\in[t-1]} \mathbb{I}\{j\in A_u, i\notin A_u\} X_{j,u} N_{ij,t-1}/N_{j,t-1},$$

is a sum of $N_{j,t-1}$ independent random variables, $N_{ij,t-1}$ of which are $(1 - N_{ij,t-1}/N_{j,t-1})^2$-sub-Gaussian and the remaining ones are $N^2_{ij,t-1}/N^2_{j,t-1}$-sub-Gaussian. So it is $N_{ij,t-1}(1 - N_{ij,t-1}/N_{j,t-1})$-sub-Gaussian, and in particular $N_{ij,t-1}$ -sub-Gaussian. ■

Appendix E. Proof of Theorem 2

Proof In the proof, we denote $\mathbf{a} \odot \mathbf{b} \triangleq (a_i b_i)_i$ the Hadamard product of two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Let $t \geq 1$, and $\delta(t) \triangleq 2(\log(t) + \log(\log(t))) + em$. Through initialization, we can assume $N_{ij,t-1} \geq 1$ for all $i, j \in [n]$ (as this only adds $n(n-1)\Delta_{\max}/2$ to the regret bound). We will decompose contributions to regret by considering the following events:

$$\mathfrak{C}_t \triangleq \left\{ \boldsymbol{\mu}^* \vee \overline{\boldsymbol{\mu}}_{t-1} \in \mathcal{C}_t(A^*) \right\},$$

$$\mathfrak{D}_t \triangleq \left\{ \mathbf{e}_{A_t}^\mathsf{T} \left(\overline{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^* \right) \leq \Delta(A_t)/2 \right\},$$

$$\mathfrak{S}_t \triangleq \left\{ \forall i, j \in [n], 0 \leq \Sigma_{ij,t} - \Sigma^*_{ij} \leq 2 g_{ij}(t) \right\}.$$

We also define

$$\widetilde{g}_{ij}(t) \triangleq 16 \left(\frac{3\log(t)}{N^2_{ij,t-1}} \vee \sqrt{\frac{3\log(t)}{N^3_{ij,t-1}}} \right) + \sqrt{\frac{48\log^2(t)}{N^4_{ij,t-1}}} + \sqrt{\frac{36\log^2(t)}{N^4_{ij,t-1}}},$$

$$\forall i \in [n],\ \Delta_{i,\min} \triangleq \min_{A \in \mathcal{A},\ i \in A,\ \Delta(A) > 0} \Delta(A),$$

$$\Delta_{i,\max} \triangleq \max_{A \in \mathcal{A},\ i \in A,\ \Delta(A) > 0} \Delta(A),$$

and

$$\forall i, j \in [n],\ \Delta_{ij,\min} \triangleq \min_{A \in \mathcal{A},\ i,j \in A,\ \Delta(A) > 0} \Delta(A),$$

$$\Delta_{ij,\max} \triangleq \max_{A \in \mathcal{A},\ i,j \in A,\ \Delta(A) > 0} \Delta(A).$$

Step 1: If $\mathfrak{C}_t, \mathfrak{D}_t$ and $\mathfrak{S}_t$ hold We have

$$\begin{aligned}
\Delta(A_t) &= (\mathbf{e}_{A^*} - \mathbf{e}_{A_t})^\mathsf{T} \boldsymbol{\mu}^* \\
&\leq \mathbf{e}_{A^*}^\mathsf{T} \boldsymbol{\mu}^* \vee \overline{\boldsymbol{\mu}}_{t-1} - \mathbf{e}_{A_t}^\mathsf{T} \boldsymbol{\mu}_t + \mathbf{e}_{A_t}^\mathsf{T} (\boldsymbol{\mu}_t - \boldsymbol{\mu}^*) \\
&\leq \mathbf{e}_{A_t}^\mathsf{T} (\boldsymbol{\mu}_t - \boldsymbol{\mu}^*) && \mathfrak{C}_t \\
&\leq \Delta(A_t)/2 + \mathbf{e}_{A_t}^\mathsf{T} \left(\boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1} \right) && \mathfrak{D}_t
\end{aligned}$$

i.e.,

$$\begin{aligned}
\Delta(A_t) &\leq 2 \mathbf{e}_{A_t}^\mathsf{T} \left(\boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1} \right) \\
&\leq 2 \sqrt{\sum_{i \in A_t} \frac{4 \left(\sum_{j \in A_t} 0 \vee \Sigma_{ij,t} + m\left(\mu_{i,t} - \overline{\mu}_{i,t-1}\right) \right) \delta(t)}{N_{i,t-1}}} && \text{Cauchy-Schwarz and } \boldsymbol{\mu}_t \in C_t(A_t) \\
&\leq 4 \sqrt{\delta(t) \sum_{i \in A_t} \left(\frac{\sum_{j \in A_t} 0 \vee \Sigma_{ij,t}}{N_{i,t-1}} + \frac{m\left(\mu_{i,t} - \overline{\mu}_{i,t-1}\right)}{\min_{j \in A_t} N_{j,t-1}} \right)}.
\end{aligned}$$

Solving the corresponding quadratic inequation in the variable $x = \mathbf{e}_{A_t}^\mathsf{T}(\boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1})$, we get

$$\begin{aligned}
\Delta(A_t) &\le 2\mathbf{e}_{A_t}^\mathsf{T}(\boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1}) \\
&\le 4\left(\sqrt{\frac{\delta(t)^2 m^2}{\min_{i\in A_t} N_{i,t-1}^2} + \sum_{i\in A_t} \frac{\delta(t)\sum_{j\in A_t} 0 \vee \Sigma_{ij,t}}{N_{i,t-1}}} + \frac{m\delta(t)}{\min_{j\in A_t} N_{j,t-1}}\right) \\
&\le 4\sqrt{\delta(t)\sum_{i\in A_t} \frac{\sum_{j\in A_t} 0 \vee \Sigma_{ij,t}}{N_{i,t-1}}} + \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \\
&\le 4\sqrt{\delta(t)\sum_{i\in A_t} \frac{\sum_{j\in A_t} 0 \vee \left(\Sigma^*_{ij} + 2g_{ij}(t)\right)}{N_{i,t-1}}} + \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \qquad \mathfrak{S}_t \\
&\le 4\sqrt{\delta(t)\sum_{i\in A_t} \frac{\sum_{j\in A_t} 0 \vee \Sigma^*_{ij}}{N_{i,t-1}}} + 4\sqrt{\delta(t)\sum_{i\in A_t} \frac{\sum_{j\in A_t} 2g_{ij}(t)}{N_{i,t-1}}} + \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \\
&\le \underbrace{4\sqrt{\delta(T)\sum_{i\in A_t} \frac{\max_{A\in\mathcal{A}, i\in A}\sum_{j\in A} 0 \vee \Sigma^*_{ij}}{N_{i,t-1}}}}_{(5)} + \underbrace{4\sqrt{\delta(T)\sum_{i,j\in A_t} \widetilde{g}_{ij}(T)}}_{(6)} + \underbrace{\frac{8m\delta(T)}{\min_{j\in A_t} N_{j,t-1}}}_{(7)}.
\end{aligned}$$

Where the last inequality uses that $N_{i,t-1} \wedge N_{j,t-1} \ge N_{ij,t-1} \forall i,j \in [n]$. From this point, we treat each term separately, using the relation

$$\mathbb{I}\{\Delta(A_t) \le (5)+(6)+(7)\} \le \mathbb{I}\{\Delta(A_t)/3 \le (5)\} + \mathbb{I}\{\Delta(A_t)/3 \le (6)\} + \mathbb{I}\{\Delta(A_t)/3 \le (7)\}.$$

We provide Theorem 4 in Appendix K, that is helpful to bound the regret on each of this 3 events. Indeed, for the first term, applying it with $\beta_{i,T} = 12^2\delta(T)\max_{A\in\mathcal{A}, i\in A}\sum_{j\in A} 0 \vee \Sigma^*_{ij}$, $\alpha_i = 1/2$, and $(I, I_t) = ([n], A_t)$ gives the bound

$$\sum_{t=1}^T \mathbb{I}\{\Delta(A_t)/3 \le (5)\}\Delta(A_t) \le 4608\log_2^2(4\sqrt{m})\sum_{i\in[n]} \frac{\delta(T)\max_{A\in\mathcal{A}, i\in A}\sum_{j\in A} 0 \vee \Sigma^*_{ij}}{\Delta_{i,\min}}.$$

The second term can be itself decomposed into two terms, bounding the max by the sum and using $\log(T) \le \delta(T)$.

$$(6) \le 4\delta(T)\sqrt{\sum_{i,j\in A_t}\left(54+\sqrt{48}\right)N_{ij,t-1}^{-2}} + 4\delta(T)^{0.75}\sqrt{16\sqrt{3}\sum_{i,j\in A_t} N_{ij,t-1}^{-1.5}}.$$

Thus, again, it is sufficient to treat each term separately. We also apply Theorem 4, but with $(I, I_t) = ([n]^2, A_t^2)$, taking respectively $\alpha_i = 1, \beta_{i,T} = 24\sqrt{54+\sqrt{48}}\delta(T)$ and $\alpha_i = 0.75, \beta_{i,T} = 192 \cdot 6^{2/3}\delta(T)$ for each term. This gives

$$\begin{aligned}
\sum_{t=1}^T \mathbb{I}\{\Delta(A_t)/3 \le (6)\}\Delta(A_t) \;&\le 1152\sqrt{6}\log_2(4m)\sum_{i,j\in[n]} \delta(T)\left(1+\log\left(\frac{\Delta_{ij,\max}}{\Delta_{ij,\min}}\right)\right) \\
&\quad + 12288\cdot 6^{2/3}\left(4^{1/3}-1\right)^{-1}\log_2(4m)\sum_{i,j\in[n]} \delta(T)\Delta_{ij,\min}^{-1/3}.
\end{aligned}$$

The last term can be analyzed in the same way by first upper bounding it as

$$(7) \leq 8m\delta(T)\sqrt{\sum_{i\in A_t}\frac{1}{N_{i,t-1}^2}}.$$

Then, taking $\alpha_i = 1, \beta_{i,T} = 24m\delta(T)$ in Theorem 4 gives

$$\sum_{t=1}^{T}\mathbb{I}\{\Delta(A_t)/3 \leq (7)\}\Delta(A_t) \leq 1152\log_2(4\sqrt{m})\sum_{i\in[n]} m\delta(T)\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right).$$

This concludes step 1; notice that all subsequent steps will aim to bound the regret by a term independent of T, over a certain event. Thus, we can see that the bounds above are the actual contributions to the rate of the regret. To show Theorem 2, we must therefore choose the regime for $\Delta \leq \Delta_{i,\min}$ so that the first term prevails over the others. In other words, we want to have

$$n^2\Big(\Delta^{-1/3} \vee (1+\log(\Delta_{\max}/\Delta))\Big) \leq c\log(m+1)\sum_{i\in[n]}\frac{\max_{A\in\mathcal{A},i\in A}\sum_{j\in A} 0 \vee \Sigma^*_{ij}}{\Delta},$$

where c is a constant. This gives exactly our condition in Theorem 2.

Step 2: If $\mathfrak{S}_t$,$\neg\mathfrak{C}_t$ hold Let $\sigma_i^2 \triangleq \sum_{j\in A^*} 0 \vee \Sigma^*_{ij}$ for all arms $i \in [n]$. We fix some $\delta \geq e \cdot m$, and define the following events:

$$\mathfrak{A}_t \triangleq \left\{\sum_{i\in A^*}\mathbb{I}\{\mu_i^* \geq \overline{\mu}_{i,t-1}\}N_{i,t-1}\frac{\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^2}{4\left(\sigma_i^2 + |A^*|\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)} \geq \delta\right\}$$

$$\forall \boldsymbol{d} \in (\mathbb{N}^*)^{A^*}, \quad \mathfrak{B}_{\boldsymbol{d},t} \triangleq \bigcap_{i\in A^*}\left\{e^{d_i-1} \leq N_{i,t-1} < e^{d_i}\right\}.$$

Notice that $\mathfrak{S}_t, \neg\mathfrak{C}_t$ implies $\mathfrak{A}_t$ for $\delta = \delta(t)$. Since each number of pulls $N_{i,t-1}$ for $i \in A^*$ is bounded by t, the number of possible $\boldsymbol{d} \in (\mathbb{N}^*)^{A^*}$ such that $\mathbb{P}[\mathfrak{B}_{\boldsymbol{d},t}] > 0$ is bounded by $\log(t)^m$. Thanks to the following Lemma 3, and an union bound on those possible $\boldsymbol{d} \in (\mathbb{N}^*)^{A^*}$, we get

$$\mathbb{P}[\mathfrak{A}_t] \leq e^{m+1}\left(\frac{(\delta-1)\log(t)}{m}\right)^m e^{-\delta},$$

so the regret under this event is bounded by a universal constant, since the upper bound above is the term of a convergent series for $\delta = \delta(t)$. Indeed, it rewrites as

$$t^{-2}e^{m+1-em}\left(\underbrace{\frac{2-\log^{-1}(t)}{m}}_{\leq 2/m} + \underbrace{2\frac{\log(\log(t))}{\log(t)} + e\log^{-1}(t)}_{\leq 2e^{e/2-1}}\right)^m,$$

that is bounded by

$$t^{-2}e\cdot\underbrace{\left(e^{1-e}\cdot 2e^{e/2-1}\right)^m}_{\leq 1}\underbrace{\left(\frac{e^{1-e/2}}{m}+1\right)^m}_{\leq e^{e^{1-e/2}}}.$$

Lemma 3 (Covering-argument) *Let* $\boldsymbol{d} \in (\mathbb{N}^*)^{A^*}$. *Then,* $\mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{\boldsymbol{d},t}] \leq \left(\frac{(\delta-1)e}{m}\right)^m e^{1-\delta}$.

Proof We rely on a covering argument. The idea is to get rid of randomness by replacing the empirical mean $\overline{\mu}_{i,t-1}$ by some non-random value x_i. Let $\boldsymbol{\zeta} \in \mathbb{R}_+^{A^*}$. For $i \in A^*$, we define $x_i(N)$ for $N \in \mathbb{R}_+$ as the unique solution $x \in (-\infty, \mu_i^*]$ of the equation $N\frac{\left(\mu_i^*-x\right)^2}{4(\sigma_i^2+|A^*|\left(\mu_i^*-x\right))} = \zeta_i$. Notice that for all $i \in A^*$, x_i is non-decreasing since $x \mapsto \frac{\left(\mu_i^*-x\right)^2}{4(\sigma_i^2+|A^*|\left(\mu_i^*-x\right))}$ is decreasing on $(-\infty, \mu_i^*]$. The event

$$\bigcap_{i\in A^*}\left\{N_{i,t-1}\frac{\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^{+2}}{4\left(\sigma_i^2 + |A^*|\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)} > \zeta_i\right\}$$

implies

$$\bigcap_{i\in A^*}\left\{\overline{\mu}_{i,t-1} \leq x_i(N_{i,t-1})\right\}.$$

Under the event $\mathfrak{B}_{\boldsymbol{d},t}$, this implies

$$\bigcap_{i\in A^*}\left\{\overline{\mu}_{i,t-1} \leq x_i(e^{d_i})\right\}. \tag{8}$$

With $\varepsilon_i \triangleq \mu_i^* - x_i(e^{d_i})$ and $\lambda_i \triangleq \frac{\varepsilon_i}{2\left(\sigma_i^2+|A^*|\varepsilon_i\right)}$, $i \in A^*$, this further implies:

$$\begin{aligned}
e^{-1}\sum_{i\in A^*}\zeta_i &= \sum_{i\in A^*} e^{d_i-1}\frac{\varepsilon_i^2}{4\left(\sigma_i^2 + |A^*|\varepsilon_i\right)} && x_i(e^{d_i}) > -\infty,\\
&\leq \sum_{i\in A^*} N_{i,t-1}\frac{\varepsilon_i^2}{4\left(\sigma_i^2 + |A^*|\varepsilon_i\right)} && \mathfrak{B}_{\boldsymbol{d},t}\\
&= \sum_{i\in A^*} N_{i,t-1}\frac{\varepsilon_i^2}{2\left(\sigma_i^2 + |A^*|\varepsilon_i\right)} - \sum_{i\in A^*} N_{i,t-1}\frac{\varepsilon_i^2}{4\left(\sigma_i^2 + |A^*|\varepsilon_i\right)}\\
&\leq \sum_{i\in A^*} N_{i,t-1}\frac{\varepsilon_i^2}{2\left(\sigma_i^2 + |A^*|\varepsilon_i\right)} - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2\frac{\varepsilon_i^2}{4\left(\sigma_i^2 + |A^*|\varepsilon_i\right)^2} && \frac{\sigma_i^2}{\sigma_i^2 + |A^*|\varepsilon_i} \leq 1,\\
&= \sum_{i\in A^*} N_{i,t-1}\lambda_i\varepsilon_i - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2\lambda_i^2\\
&\leq \sum_{i\in A^*} N_{i,t-1}\lambda_i\left(\mu_i^* - \overline{\mu}_{i,t-1}\right) - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2\lambda_i^2 && \text{using (8)},\\
&= \sum_{u\in[t-1]}\left((\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})^\intercal(\boldsymbol{\mu}^* - \mathbf{X}_u) - (\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})^\intercal\mathbf{D}(\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})\right),
\end{aligned}$$

where $\mathbf{D}$ is the diagonal matrix with $D_{ii} = \sigma_i^2$ for all $i \in [n]$. For all $u \in [t-1]$, since $\boldsymbol{\lambda} \geq 0$, we can write the following axis-realignment inequality

$$\begin{aligned}
(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} \boldsymbol{\Sigma}^* (\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*}) &= \sum_{i \in A_u \cap A^*} \sum_{j \in A_u \cap A^*} \Sigma^*_{ij} \lambda_i \lambda_j \\
&\leq \sum_{i \in A_u \cap A^*} \sum_{j \in A_u \cap A^*} \frac{0 \vee \Sigma^*_{ij}}{2} \big(\lambda_i^2 + \lambda_j^2\big) \\
&= \sum_{i \in A_u \cap A^*} \left(\sum_{j \in A_u \cap A^*} 0 \vee \Sigma^*_{ij} \right) \lambda_i^2 \\
&\leq (\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} \mathbf{D} (\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*}).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
e^{-1} \sum_{i \in A^*} \zeta_i &\leq \sum_{u \in [t-1]} \big((\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u) - (\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} \boldsymbol{\Sigma}^* (\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*}) \big) \\
&\leq \sum_{u \in [t-1]} \left((\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u) - \log \mathbb{E} \left[e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right] \right),
\end{aligned}$$

where the last inequality uses Assumption 2 and $\|\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*}\|_1 \leq 1/2$. Now, notice that

$$\mathbb{E} \left[\exp \left(\sum_{u \in [t-1]} \left((\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u) - \log \mathbb{E} \left[e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right] \right) \right) \right]$$

equals

$$\begin{aligned}
\mathbb{E} \left[\prod_{u \in [t-1]} \frac{e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u)}}{\mathbb{E} \left[e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right]} \right] &= \prod_{u \in [t-1]} \mathbb{E} \left[\frac{e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u)}}{\mathbb{E} \left[e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right]} \right] \\
&= 1,
\end{aligned}$$

so from Markov inequality, we get the following bound:

$$\mathbb{P} \left[\sum_{u \in [t-1]} \left((\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_u) - \log \mathbb{E} \left[e^{(\boldsymbol{\lambda} \odot \mathbf{e}_{A_u \cap A^*})^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right] \right) \geq e^{-1} \sum_{i \in A^*} \zeta_i \right] \leq e^{-\sum_{i \in A^*} \zeta_i e^{-1}},$$

thus, we showed that

$$\mathbb{P} \left[\mathfrak{B}_{\boldsymbol{d},t} \cap \bigcap_{i \in A^*} \left\{ N_{i,t-1} \frac{\big(\mu_i^* - \overline{\mu}_{i,t-1}\big)^{+2}}{4\big(\sigma_i^2 + |A^*| \big(\mu_i^* - \overline{\mu}_{i,t-1}\big)\big)} > \zeta_i \right\} \right] \leq e^{-\sum_{i \in A^*} \zeta_i e^{-1}},$$

i.e.

$$\mathbb{P} \left[\bigcap_{i \in A^*} \left\{ \mathbb{I}\{\mathfrak{B}_{\boldsymbol{d},t}\} N_{i,t-1} \frac{\big(\mu_i^* - \overline{\mu}_{i,t-1}\big)^{+2}}{4\big(\sigma_i^2 + |A^*| \big(\mu_i^* - \overline{\mu}_{i,t-1}\big)\big)} > \zeta_i \right\} \right] \leq e^{-\sum_{i \in A^*} \zeta_i e^{-1}},$$

By Lemma 8 of Magureanu et al. (2014), since $\delta \geq em$, we have

$$\begin{aligned}\mathbb{P}[\mathfrak{B}_{\boldsymbol{d},t} \cap \mathfrak{A}_t] &= \mathbb{P}\left[\mathfrak{B}_{\boldsymbol{d},t} \cap \left\{\sum_{i\in A^*} N_{i,t-1}\frac{\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^{+2}}{4\left(\sigma_i^2 + |A^*|\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)} \geq \delta\right\}\right] \\ &= \mathbb{P}\left[\sum_{i\in A^*} \mathbb{I}\{\mathfrak{B}_{\boldsymbol{d},t}\} N_{i,t-1}\frac{\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^{+2}}{4\left(\sigma_i^2 + |A^*|\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)} \geq \delta\right] \\ &\leq \left(\frac{(\delta-1)e}{m}\right)^m e^{1-\delta}.\end{aligned}$$

■

Step 3: If $\neg\mathfrak{D}_t$ hold The regret under this event can be bounded by $8nm^2\Delta_{\max}/\Delta^2$ using exactly the same method as Lemma 2 of Degenne and Perchet (2016).

Step 4: If $\neg\mathfrak{S}_t$ hold From Proposition 2, the regret under this event is bounded by a universal constant.

Putting it all together Finally, we have shown that there exists two universal constant c, c' satisfying the following (we display the scaled back (by κ) version of the regret bound to get the dependence into κ)

$$\begin{aligned}R_T \leq \Delta_{\max}\left(\frac{n(n-1)}{2} + \frac{8nm^2}{\Delta^2} + c\right) + c'\log(m+1)\delta(T)\Bigg[&\log(m+1)\sum_{i\in[n]}\frac{\max_{A\in\mathcal{A}, i\in A}\sum_{j\in A} 0 \vee \Sigma^*_{ij}}{\Delta_{i,\min}} \\ &+ \sum_{i,j\in[n]}\kappa\left(1 + \log\left(\frac{\Delta_{ij,\max}}{\Delta_{ij,\min}}\right)\right) \\ &+ \sum_{i\in[n]} m\kappa\left(1 + \log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right) \\ &+ \sum_{i,j\in[n]}\frac{\kappa^{4/3}}{\Delta_{ij,\min}^{1/3}}\Bigg]. \qquad (9)\end{aligned}$$

■

Appendix F. The bound of Corollary 1

The corollary is obtained in the same way as Theorem 2. We can underline the difference that we don't have to construct n^2 covariance estimates, but only n (only the ν_i^*'s). On the other hand, as the estimation uses sub-Gaussian variables, we don't use sub-exponential concentration, which

removes one term from the previous result. The obtained bound is

$$R_T \leq \Delta_{\max}\left(n + \frac{8nm^2}{\Delta^2} + c\right) + c' \log(m+1)\delta(T)\Bigg[\log(m+1)\sum_{i\in[n]}\frac{\nu_i^*(s\wedge m)}{\Delta_{i,\min}} + \sum_{i\in[n]} m\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right) + \sum_{i\in[n]}\frac{(s\wedge m)^{2/3}}{\Delta_{i,\min}^{1/3}}\Bigg], \tag{10}$$

where c and c' are two constants. Notice that to make the first term dominates the others, we must have

$$n(s\wedge m)^{2/3}/\Delta^{1/3} \vee (nm(1+\log(\Delta_{\max}/\Delta))) \leq c\log(m+1)\sum_{i\in[n]}\frac{\nu_i^*(s\wedge m)}{\Delta},$$

for some constant c, which gives our condition in Corollary 1.

Appendix G. Proof of Theorem 3

Proof Consider $\mathcal{A}$ containing n/m disjoint actions $A_1, \ldots, A_{n/m}$ composed of m arms. $\mathbf{X}$ is constructed as follows: $(1 \vee s/m)$ different actions are randomly chosen among $\mathcal{A}$, with equal probability, except the one for action A_1, that have an offset of δ. From

$$(1\vee s/m) = \mathbb{E}\left[\sum_{A\in\mathcal{A}}\mathbb{I}\{A \text{ is chosen}\}\right] = (n/m-1)(\mathbb{P}[A_1 \text{ is chosen}]-\delta) + \mathbb{P}[A_1 \text{ is chosen}],$$

we have $\mathbb{P}[A_1 \text{ is chosen}] = (1\vee s/m)m/n + \delta(1-m/n)$. We pose $X_i = 1$ for i spanning the $(s\wedge m)$ first arm of each chosen action (the other components are set to 0). Remark that $\mathbf{X}$ is $s-$sparse with this construction.

This problem reduces to a standard bandit problem with n/m Bernoulli arms. However, we have an additional piece of information, namely that the sum of the means is s. Thus, we can't apply the lower bound from Lai and Robbins (1985), since the distribution family has not a product form (changing the mean of one arm, we have to make sure that the sum of the means doesn't change, so we have to change at least another mean). Instead, we use the lower bound result from Graves and Lai (1997), where we can increase the mean of one arm i while decreasing the mean of the others. Scaling the regret by $(s\wedge m)^{-1}$, we want to upper bound

$$\mathrm{KL}\left(\mathbb{P}_{\frac{1}{(s\wedge m)}\sum_{i\in A_k}X_i} \| \mathbb{P}_{\frac{1}{(s\wedge m)}\sum_{i\in A_1}X_i}\right) = \mathrm{kl}\left(\left(\frac{m}{n}\vee\frac{s}{n}\right)-\delta\frac{m}{n}, \left(\frac{m}{n}\vee\frac{s}{n}\right)+\delta\left(1-\frac{m}{n}\right)\right),$$

which corresponds to an arm i that becomes a best arm for the new distribution. We also want to upper bound

$$\mathrm{kl}\left(\left(\frac{m}{n}\vee\frac{s}{n}\right)-\delta\frac{m}{n}-\frac{\delta}{\frac{n}{m}-2}, \left(\frac{m}{n}\vee\frac{s}{n}\right)-\delta\frac{m}{n}\right),$$

which corresponds to the decrease of the mean of each sub-optimal arm k different from i (so that the sum of the mean remain constant). We are going to use the inequality $\mathrm{kl}(x,y) \leq \frac{(x-y)^2}{y(1-y)}$ for all $x, y \in (0,1)$. Since $\frac{ms}{2(n-m)} \geq \Delta = (s \wedge m)\delta$, we have $\delta\frac{m}{n} \leq \delta(1-\frac{m}{n}) \leq \big(\frac{m}{n} \vee \frac{s}{n}\big)/2 \leq 1/4$, and thus

$$\Big(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n}\Big)\Big(1 - \Big(\frac{m}{n} \vee \frac{s}{n}\Big) + \delta\frac{m}{n}\Big) \geq \Big(\frac{m}{n} \vee \frac{s}{n}\Big)/4,$$
$$\Big(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) + \delta\Big(1 - \frac{m}{n}\Big)\Big)\Big(1 - \Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\Big(1 - \frac{m}{n}\Big)\Big) \geq \Big(\frac{m}{n} \vee \frac{s}{n}\Big)/4.$$

Thus, we get the upper bounds

$$\mathrm{kl}\Big(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n}, \Big(\frac{m}{n} \vee \frac{s}{n}\Big) + \delta\Big(1 - \frac{m}{n}\Big)\Big) \leq \frac{4\delta^2}{\big(\frac{m}{n} \vee \frac{s}{n}\big)} \tag{11}$$

$$\mathrm{kl}\left(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n} - \frac{\delta}{\frac{n}{m} - 2}, \Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n}\right) \leq \frac{4\delta^2}{\big(\frac{n}{m} - 2\big)^2\big(\frac{m}{n} \vee \frac{s}{n}\big)}. \tag{12}$$

From Graves and Lai (1997), we have the lower bound

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq (s \wedge m) \inf_{\mathbf{c}} \sum_{k=2}^{n/m} \delta c_k,$$

where the above infimum is over all $c_2, \ldots, c_{n/m}$ in $\mathbb{R}_+$ such that for all $i \in \{2, \ldots, n/m\}$,

$$c_i \mathrm{kl}\Big(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n}, \Big(\frac{m}{n} \vee \frac{s}{n}\Big) + \delta\Big(1 - \frac{m}{n}\Big)\Big) + \sum_{k=2, k\neq i}^{n/m} c_k \mathrm{kl}\left(\Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n} - \frac{\delta}{\frac{n}{m} - 2}, \Big(\frac{m}{n} \vee \frac{s}{n}\Big) - \delta\frac{m}{n}\right) \geq 1.$$

Using the bounds (11) and (12), we can relax the above constraint as

$$\forall i \in \{2, \ldots, n/m\},\ c_i \frac{4\delta^2}{\big(\frac{m}{n} \vee \frac{s}{n}\big)} + \sum_{k=2, k\neq i}^{n/m} c_k \frac{4\delta^2}{\big(\frac{n}{m} - 2\big)^2\big(\frac{m}{n} \vee \frac{s}{n}\big)} \geq 1.$$

By symmetry of the constraint with respect to c_i, and by linearity of the objective, there is a maximizer $\mathbf{c}$ that satisfies $c_1 = \cdots = c_{n/m} = c$, with

$$4c\delta^2\left(\frac{1}{\big(\frac{m}{n} \vee \frac{s}{n}\big)} + \frac{1}{\big(\frac{n}{m} - 2\big)\big(\frac{m}{n} \vee \frac{s}{n}\big)}\right) = 1.$$

Thus, since $\Delta = (s \wedge m)\delta$, we get

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq \frac{s(s \wedge m)(1 - 2m/n)}{4\Delta}.$$

Notice that we recover the full information case (with a lower bound that equals 0) when $n/m = 2$, as expected. ■

Appendix H. Proof of Lemma 1

Proof We use the fact that $\sum_{j\in A}|X_j| \leq (s \wedge m)$. This gives

$$\begin{aligned}\sum_{j\in A} 0 \vee \Sigma^*_{ij} &= \sum_{j\in A} 0 \vee \mathbb{E}\big[X_iX_j - \mu^*_i\mu^*_j\big] \\ &\leq \sum_{j\in A}\big(\mathbb{E}[|X_iX_j|] + \big|\mu^*_i\mu^*_j\big|\big) \\ &= \mathbb{E}\left[|X_i|\sum_{j\in A}|X_j|\right] + |\mu^*_i|\sum_{j\in A}\big|\mu^*_j\big| \\ &\leq 2\mathbb{E}[|X_i|](s\wedge m).\end{aligned}$$

■

Appendix I. Confidence regions comparison

We give here the two algorithms CUCB-V and CUCB-KL, which, as we have seen, also matches the lower bound given to the Theorem 3, in the specific regime where $s \geq m$. Both the two algorithms rely on the same optimization $A_t = \arg\max_{A\in\mathcal{A}} \mathbf{e}_A^\intercal \boldsymbol{\mu}_t$, where the vector $\boldsymbol{\mu}_t$ is defined for CUCB-V as

$$\forall i \in [n], \quad \mu_{i,t} \triangleq 1 \wedge \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{2\zeta\overline{\sigma}^2_{i,t-1}\log(t)}{N_{i,t-1}}} + \frac{3\zeta\log(t)}{N_{i,t-1}}\right),$$

where

$$\overline{\sigma}^2_{i,t-1} \triangleq \frac{\sum_{t'\in[t-1]}\mathbb{I}\{i=i_{t'}\}\big(X_{i,t'} - \overline{\mu}_{i,t-1}\big)^2}{N_{i,t-1}},$$

and for CUCB-KL as

$$\forall i \in [n], \quad \mu_{i,t} \text{ is the unique solution } x \text{ to } N_{i,t-1}\mathrm{kl}\big(\overline{\mu}_{i,t-1}, x\big) = \zeta\log(t) \text{ such that } x \in [\overline{\mu}_{i,t-1}, 1].$$

We take $\zeta = 1.2$ (although all $\zeta > 1$ are valid). The algorithms above can also be seen as a bilinear maximization where $\boldsymbol{\mu}_t$ is maximized over a confidence region that is a Cartesian product one 1-demendional confidence intervals. We illustrate in Figure 2 the difference between the confidence region considered in ESCB-C (when the correlation is low) and CUCB-KL. The red points represent $\boldsymbol{\mu}_t$ for each region. It can be seen that the Cartesian product confidence region greatly overestimates the risk in directions that are not close to the axes, giving rise to over-exploration. It is important to note however that this price to pay can be interesting in practice, because the corresponding algorithms are then very efficient (LP over $\mathcal{A}$, supposed possible[5]). As we noted in Remark 1, considering the intersection between the two confidence regions gives rise to an even tighter region, and therefore a better regret bound.

5. Otherwise an approximation regret would be a more appropriate performance measure to consider.

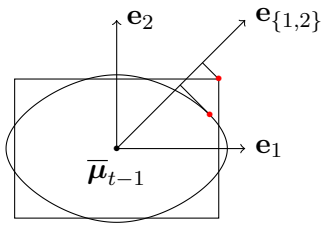

Figure 2: *Confidence regions build by* ESCB-C *(the pseudo-ellipse), and* CUCB-KL *(the rectangle), for $\|\cdot\|_1$ constrained outcomes. Notice that* CUCB-KL *has slightly better confidence intervals along the axis, but that* ESCB-C *is better in the direction $\mathbf{e}_{\{1,2\}}$.*

Appendix J. Implementation using the Lovász extension

From the step 1 of the proof of Theorem 2, we have that

$$\mathbf{e}_{A_t}^\mathsf{T}\boldsymbol{\mu}_t \leq \mathbf{e}_{A_t}^\mathsf{T}\overline{\boldsymbol{\mu}}_{t-1} + 2\sqrt{\delta(t)\sum_{i\in A_t}\frac{\sum_{j\in A_t} 0 \vee \Sigma_{ij,t}}{N_{i,t-1}}} + 4m\delta(T)\sqrt{\sum_{i\in A_t}\frac{1}{N_{i,t-1}^2}}.$$

Since the final bound of Theorem 2 relies on the above upper bound, in Algorithm 1, instead of maximizing $A \mapsto \max_{\boldsymbol{\mu}\in\mathcal{C}_t(A)} \mathbf{e}_A^\mathsf{T}\boldsymbol{\mu}$, we can maximize

$$A \mapsto \mathbf{e}_A^\mathsf{T}\overline{\boldsymbol{\mu}}_{t-1} + 2\sqrt{\delta(t)\sum_{i\in A}\frac{\sum_{j\in A} 0 \vee \Sigma_{ij,t}}{N_{i,t-1}}} + 4m\delta(T)\sqrt{\sum_{i\in A}\frac{1}{N_{i,t-1}^2}}.$$

Our goal here is to provide a continuous extension of the above set function that is concave on $[0,1]^n$, and thus efficient to maximize. The linear term trivially extends to the linear function $\mathbf{x} \mapsto \mathbf{x}^\mathsf{T}\overline{\boldsymbol{\mu}}_{t-1}$. The last two term can be extended relying on the Lovász extension (Lovász, 1983). We recall that the Lovász extension of a set function f is defined as $f^L(\mathbf{x}) \triangleq \mathbb{E}[f(\{i \in [n],\ x_i \geq U\})]$, where the expectation is over $U \sim \mathcal{U}[0,1]$. The Lovász extension is concave if and only if f is a supermodular function (Lovász, 1983), i.e.,

$$f(A) + f(B) \leq f(A \cup B) + f(A \cap B) \quad \forall A, B \subset [n].$$

It is easy to check that a function $G : A \mapsto \sum_{i\in A}\sum_{j\in A} a_{ij}$ is supermodular for $a_{ij} \geq 0$, so its Lovász extension is concave. Composing by the square root, we thus have a concave extension of the second and last term.

After the maximization of the extension, a continuous maximizer $\mathbf{x}_t$ is returned, and the agent plays $A_t = \{i \in [n],\ x_{i,t} \geq U\}$ where $U \sim \mathcal{U}[0,1]$. The analysis holds the same, except in Proposition 3, where counters are updated only for the chosen set. Let σ_t be a permutation such that $x_{\sigma_t(1),t} \geq \ldots, \geq x_{\sigma_t(n),t}$. Then, the set $S_j = \{\sigma_t(1), \ldots, \sigma_t(j)\}$ is chosen with probability $p_{j,t} = x_{\sigma_t(j)} - x_{\sigma_t(j+1)}$ (with the convention $x_{\sigma_t(n+1)} = 0$). The continuous extension evaluated at $\mathbf{x}_t$ is of the form

$$\sum_j p_{j,t}\mathbf{e}_{S_j}^\mathsf{T}\overline{\boldsymbol{\mu}}_{t-1} + \sqrt{\sum_j p_{j,t}G_1(S_j)} + \sqrt{\sum_j p_{j,t}G_2(S_j)},$$

where G_1 and G_2 are the supermodular functions corresponding to the second and last term respectively. Since the probabilities $p_{j,t}$ are inside the square root, applying the *Probabilistically triggered arms* setting of Wang and Chen (2017a) gives an extra factor of $1 + \log\left(\frac{m\log(T)}{\Delta^2}\right)$.

Appendix K. General stochastic combinatorial semi-bandits results

Theorem 4 (Regret bound for ℓ_2-norm error) *Let I be a set of index. For all $i \in I$, let $(\alpha_i, \beta_{i,T}) \in [1/2, 1) \times \mathbb{R}_+$. Let I_t be a subset of I such that for all $i \in I_t$, $N_{i,t} = N_{i,t-1} + 1$. We pose $\Delta_t \triangleq \Delta(A_t)$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \left\| \sum_{i \in I_t} \frac{\beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \right\}.$$

Then,

$$\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\} \Delta_t \leq 4 \log_2(4\sqrt{m}) \sum_{i \in I} \beta_{i,T} \eta_i,$$

where

$$\eta_i \triangleq \begin{cases} 8 \log_2(4\sqrt{m}) \Delta_{i,\min}^{-1} & \text{if } \alpha_i = 1/2 \\ \left(\left(2^{-\frac{1}{\alpha_i}} - 2^{-2}\right)(1-\alpha_i) \Delta_{i,\min}^{\frac{1-\alpha_i}{\alpha_i}} \right)^{-1} & \text{if } 1/2 < \alpha_i < 1 \\ 4\left(1 + \log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right) & \text{if } \alpha_i = 1. \end{cases}$$

Proof Let $t \geq 1$. We define $\Lambda_t \triangleq \left\| \sum_{i \in I_t} \beta_{i,T}^{\alpha_i} N_{i,t-1}^{-\alpha_i} \mathbf{e}_i \right\|_2$. We start by a simple lower bound on Λ_t, holding for any $j \in I_t$,

$$\Lambda_t \geq \left\| \frac{\beta_{j,T}^{\alpha_j} \mathbf{e}_j}{N_{j,t}^{\alpha_j}} \right\|_2 = \frac{\beta_{j,T}^{\alpha_j}}{N_{j,t}^{\alpha_j}}. \tag{13}$$

We then use the same reverse amortisation technique than in Wang and Chen (2017b).

$$\begin{aligned}
\Lambda_t &= -\Lambda_t + \left\| \sum_{i \in I_t} \frac{2\beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\
&= -\left\| \frac{\Lambda_t \mathbf{e}_{I_t}}{\|\mathbf{e}_{I_t}\|_2} \right\|_2 + \left\| \sum_{i \in I_t} \frac{2\beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\
&\leq \left\| \sum_{i \in I_t} \left(\frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda_t}{\|\mathbf{e}_{I_t}\|_2} \right)^+ \mathbf{e}_i \right\|_2 \\
&= \left\| \sum_{i \in I_t} \left(\frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda_t}{\|\mathbf{e}_{I_t}\|_2} \right)^+ \mathbb{I}\left\{ \Lambda_t \geq \frac{\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \right\} \mathbf{e}_i \right\|_2 \qquad \text{Using (13)} \\
&\leq \left\| \sum_{i \in I_t} \mathbb{I}\left\{ 2\Lambda_t \geq \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda_t}{\|\mathbf{e}_{I_t}\|_2} \right\} \frac{2\beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2.
\end{aligned}$$

We now decompose the interval $[2, 1/\|\mathbf{e}_{I_t}\|_2]$ using a peeling:

$$[2, 1/\|\mathbf{e}_{I_t}\|_2] \subset \bigcup_{k=0}^{\lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil} [2^{1-k}, 2^{-k}].$$

This induces a partition of the set of indices:

$$\mathbb{I}\left\{ i \in I_t,\ 2\Lambda_t \geq \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda_t}{\|\mathbf{e}_{I_t}\|_2} \right\} \subset \bigcup_{k=0}^{\lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil} J_{k,t},$$

where for all interger $1 \leq k \leq \lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil$,

$$J_{k,t} \triangleq \left\{ i \in I_t,\ 2^{1-k}\Lambda_t \geq \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq 2^{-k}\Lambda_t \right\}.$$

We can thus upper bound Λ_t^2 using this decomposition

$$\begin{aligned}
\Lambda_t^2 &\leq \left\| \sum_{i \in I_t} \mathbb{I}\left\{ 2\Lambda_t \geq \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda_t}{\|\mathbf{e}_{I_t}\|_2} \right\} \frac{2\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2^2 \\
&\leq \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil} \left\| \sum_{i \in J_{k,t}} \frac{2\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2^2 \\
&\leq \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil} 2^{2-2k}\Lambda_t^2 \left\|\mathbf{e}_{J_{k,t}}\right\|_2^2.
\end{aligned}$$

This last inequality implies that there must exist one integer k_t such that $|J_{k_t,t}| = \left\|\mathbf{e}_{J_{k_t,t}}\right\|_2^2 \geq 2^{2k_t-2}(1 + \lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil)^{-1}$. We now upper bound $\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t$, using $|I_t| \leq m$, i.e.,

$$\lceil \log_2(\|\mathbf{e}_{I_t}\|_2)\rceil \leq \lceil \log_2(m)/2\rceil.$$

$$\begin{aligned}
\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t &\leq \sum_{t=1}^T \sum_{k=0}^{\lceil \log_2(m)/2\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\}\Delta_t \\
&\leq \sum_{t=1}^T \sum_{k=0}^{\lceil \log_2(m)/2\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\} \sum_{i \in I} \mathbb{I}\{i \in J_{k,t}\}\Delta_t 2^{2-2k}(\lceil \log_2(m)/2\rceil + 1) \\
&\leq \sum_{t=1}^T \sum_{k=0}^{\lceil \log_2(m)/2\rceil} \sum_{i \in I} \mathbb{I}\left\{ i \in I_t,\ N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\Delta_t} \right\} \Delta_t 2^{2-2k}(\lceil \log_2(m)/2\rceil + 1) \\
&= (\lceil \log_2(m)/2\rceil + 1) \sum_{k=0}^{\lceil \log_2(m)/2\rceil} 2^{2-2k} \sum_{i \in I} \underbrace{\sum_{t=1}^T \mathbb{I}\left\{ i \in I_t,\ N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\Delta_t} \right\} \Delta_t}_{(14)_{i,k}}.
\end{aligned}$$

Applying Proposition 3 gives

$$(14)_{i,k} \leq \mathbb{I}\{\alpha_i < 1\}\frac{\beta_{i,T}2^{\frac{k+1}{\alpha_i}}}{1-\alpha_i}\Delta_{i,\min}^{1-1/\alpha_i} + \mathbb{I}\{\alpha_i = 1\}2^{k+1}\beta_{i,T}\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right).$$

So we get, using $\lceil \log_2(m)/2 \rceil + 1 \leq \log_2(4\sqrt{m})$,

$$\sum_{t=1}^{T}\mathbb{I}\{\mathfrak{A}_t\}\Delta_t \leq 4\log_2(4\sqrt{m})\sum_{i\in I}\beta_{i,T}\eta_i,$$

$$\text{with } \eta_i = \begin{cases} 8\log_2(4\sqrt{m})\Delta_{i,\min}^{-1} & \text{if } \alpha_i = 1/2 \\ \left(\left(2^{-\frac{1}{\alpha_i}} - 2^{-2}\right)(1-\alpha_i)\Delta_{i,\min}^{\frac{1-\alpha_i}{\alpha_i}}\right)^{-1} & \text{if } 1/2 < \alpha_i < 1 \\ 4\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right) & \text{if } \alpha_i = 1. \end{cases}$$

■

Proposition 3 *Let $i \in I$ and $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing function, integrable on $[\Delta_{i,\min}, \Delta_{i,\max}]$. Then*

$$\sum_{t=1}^{T}\mathbb{I}\{i \in I_t,\ N_{i,t-1} \leq f_i(\Delta_t)\}\Delta_t \leq f_i(\Delta_{i,\min})\Delta_{i,\min} + \int_{\Delta_{i,\min}}^{\Delta_{i,\max}} f_i(x)\mathrm{d}x.$$

In particular,

- *If $f_i(x) = \beta_{i,T}x^{-1/\alpha_i}$, $\alpha_i \in (0,1)$ and $\beta_{i,T} \geq 0$, then*

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{I}\{i \in I_t,\ N_{i,t-1} \leq f_i(\Delta_t)\}\Delta_t &\leq \Delta_{i,\min}^{1-1/\alpha_i}\frac{\beta_{i,T}}{1-\alpha_i} - \Delta_{i,\max}^{1-1/\alpha_i}\frac{\alpha_i\beta_{i,T}}{1-\alpha_i} \\ &\leq \Delta_{i,\min}^{1-1/\alpha_i}\frac{\beta_{i,T}}{1-\alpha_i}.\end{aligned}$$

- *If $f_i(x) = \beta_{i,T}x^{-1}$, $\beta_{i,T} \geq 0$, then*

$$\sum_{t=1}^{T}\mathbb{I}\{i \in I_t,\ N_{i,t-1} \leq f_i(\Delta_t)\}\Delta_t \leq \beta_{i,T}\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right).$$

Proof Consider $\Delta_{i,\max} = \Delta_{i,1} \geq \Delta_{i,2} \geq \cdots \geq \Delta_{i,K_i} = \Delta_{i,\min}$ being all possible values for Δ_t when $i \in I_t$. We define a dummy gap $\Delta_{i,0} = \infty$ and let $f_i(\Delta_{i,0}) = 0$. In (15), we first break the range $(0, f_i(\Delta_t)]$ of the counter $N_{i,t-1}$ into sub intervals:

$$(0, f_i(\Delta_t)] = (f_i(\Delta_{i,0}), f_i(\Delta_{i,1})] \cup \cdots \cup (f_i(\Delta_{i,k_t-1}), f_i(\Delta_{i,k_t})],$$

where k_t is the index such that $\Delta_{i,k_t} = \Delta_t$. This index k_t exists by assumption that the subdivision contains all possible values for Δ_t when $i \in I_t$. Notice that in (15), we do not explicitly use k_t, but instead sum over all $k \in [K_i]$ and filter against the event $\{\Delta_{i,k} \geq \Delta_t\}$, which is equivalent to summing over $k \in [k_t]$.

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{I}\{i \in I_t,\ N_{i,t-1} \leq f_i(\Delta_t)\}\Delta_t \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in I_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k}), \Delta_{i,k} \geq \Delta_t\}\Delta_t. \end{aligned} \tag{15}$$

Over each event that $N_{i,t-1}$ belongs to the interval $(f_i(\Delta_{i,k-1}), f_i(\Delta_{i,k})]$, we upper bound the suffered gap Δ_t by $\Delta_{i,k}$.

$$(15) \leq \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in I_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k}), \Delta_{i,k} \geq \Delta_t\}\Delta_{i,k}. \tag{16}$$

Then, we further upper bound the summation by adding events that $N_{i,t-1}$ belongs to the remaining intervals $(f_i(\Delta_{i,k-1}), f_i(\Delta_{i,k})]$ for $k_t < k \leq K_i$, associating them to a suffered gap $\Delta_{i,k}$. This is equivalent to removing the filtering against the event $\{\Delta_{i,k} \geq \Delta_t\}$.

$$(16) \leq \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in I_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k})\}\Delta_{i,k}. \tag{17}$$

Now, we invert the summation over t and the one over k.

$$(17) = \sum_{k=1}^{K_i}\sum_{t=1}^{T} \mathbb{I}\{i \in I_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k})\}\Delta_{i,k}. \tag{18}$$

For each $k \in [K_i]$, the number of times $t \in [T]$ that the counter $N_{i,t-1}$ belongs to $(f_i(\Delta_{i,k-1}), f_i(\Delta_{i,k})]$ can be upper bounded by the number of integers in this interval. This is due to the event $\{i \in I_t\}$, imposing that $N_{i,t-1}$ is incremented, so $N_{i,t-1}$ cannot be worth the same integer for two different times t satisfying $i \in I_t$. We use the fact that for all $x, y \in \mathbb{R}$, $x \leq y$, the number of integers in the interval $(x, y]$ is exactly $\lfloor y \rfloor - \lfloor x \rfloor$.

$$(18) \leq \sum_{k=1}^{K_i} (\lfloor f_i(\Delta_{i,k}) \rfloor - \lfloor f_i(\Delta_{i,k-1}) \rfloor)\Delta_{i,k}. \tag{19}$$

We then simply expand the summation, and some terms are cancelled (remember that $f_i(\Delta_{i,0}) = 0$).

$$(19) = \lfloor f_i(\Delta_{i,K_i}) \rfloor \Delta_{i,K_i} + \sum_{k=1}^{K_i-1} \lfloor f_i(\Delta_{i,k}) \rfloor (\Delta_{i,k} - \Delta_{i,k+1}) \tag{20}$$

We use $\lfloor x \rfloor \leq x$ for all $x \in \mathbb{R}$. Finally, we recognize a right Riemann sum, and use the fact that f_i is non increasing to upper bound each $f_i(\Delta_{i,k})(\Delta_{i,k} - \Delta_{i,k+1})$ by $\int_{\Delta_{i,k+1}}^{\Delta_{i,k}} f_i(x)\mathrm{d}x$, for all $k \in [K_i - 1]$.

$$(20) \leq f_i(\Delta_{i,K_i})\Delta_{i,K_i} + \sum_{k=1}^{K_i-1} f_i(\Delta_{i,k})(\Delta_{i,k} - \Delta_{i,k+1}) \tag{21}$$

$$\leq f_i(\Delta_{i,K_i})\Delta_{i,K_i} + \int_{\Delta_{i,K_i}}^{\Delta_{i,1}} f_i(x)\mathrm{d}x. \tag{22}$$

■