Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits

Pierre Perrault
Inria Lille — ENS Paris-Saclay
pierre.perrault@inria.fr

Etienne Boursier
ENS Paris-Saclay
etienne.boursier1@gmail.com

Vianney Perchet
ENSAE — Criteo AI Lab
vianney.perchet@normalesup.org

Michal Valko
DeepMind Paris — Inria Lille
valkom@deepmind.com

Abstract

We investigate stochastic combinatorial multi-armed bandit with semi-bandit feedback (CMAB). In CMAB, the question of the existence of an efficient policy with an optimal asymptotic regret (up to a factor poly-logarithmic with the action size) is still open for many families of distributions, including mutually independent outcomes, and more generally the multivariate *sub-Gaussian* family. We propose to answer the above question for these two families by analyzing variants of the *Combinatorial Thompson Sampling* policy (CTS). For mutually independent outcomes in $[0, 1]$, we propose a tight analysis of CTS using Beta priors. We then look at the more general setting of multivariate sub-Gaussian outcomes and propose a tight analysis of CTS using Gaussian priors. This last result gives us an alternative to the *Efficient Sampling for Combinatorial Bandit* policy (ESCB), which, although optimal, is not computationally efficient.

1 Introduction

Stochastic multi-armed bandits (MAB) Robbins [1952], Berry and Fristedt [1985], Lai and Robbins [1985] are decision-making frameworks in which a learning *agent* acts sequentially in an uncertain environment. At every round $t \in \mathbb{N}^*$, the agent must select one arm from a pool of n arms, denoted by $[n] \triangleq \{1, \dots, n\}$, using a learning *policy* based on the feedback collected from the previous rounds. Then it obtains as feedback a reward (also called *outcome*) $X_{i,t} \in \mathbb{R}$ — a random variable sampled from $\mathbb{P}_{X_i}$, independently from previous rounds — where i is the selected arm and $\mathbb{P}_{X_i}$ is a probability distribution — unknown to the agent — of mean μ_i^*. The goal for the agent is to maximize the cumulative reward over a total of T rounds (T may be unknown[1]). The performance metric of a policy is the regret, i.e., the expectation of the difference over T rounds of the cumulative reward between the policy that always picked the arm with the highest expected reward and the learning policy. MAB models the classical dilemma between exploration and exploitation, i.e., whether to continue exploring arms to obtain more information (and thus strengthen the confidence in the estimates of the distributions $\mathbb{P}_{X_i}$), or to use the information gathered by playing the best arm according to the observations so far.

In this paper, we study stochastic combinatorial multi-armed bandit (CMAB) [Cesa-Bianchi and Lugosi, 2012], which is an extension of MAB where the agent selects a *super arm* (or *action*) $A_t \in \mathcal{A} \subset \mathcal{P}([n])$ at each round t. The set $\mathcal{A}$ is the *action space*, defined as a collection of subsets

[1]We recall here the fact that in MAB, whether the horizon T is known or not is not really relevant as algorithms can be easily adapted [Degenne and Perchet, 2016a].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

of the (base) arms. The kind of reward and feedback varies depending on the problem at hand. We consider the *semi-bandit* setting, where the feedback includes the outcomes of all base arms in the played super arm. Formally, the agent observes[2] $\mathbf{X}_t \odot \mathbf{e}_{A_t} \triangleq (X_{i,t}\mathbb{I}\{i \in A_t\})_{i\in[n]}$ and the reward, given the choice of A_t, is a function of $\boldsymbol{\mu}^* \odot \mathbf{e}_{A_t}$ (traditionally, the reward is linear and equal to $\mathbf{e}_{A_t}^\intercal \boldsymbol{\mu}^*$, but our analysis goes beyond this setting). In recent years, CMAB has attracted a lot of interest (see e.g. Gai et al. [2012], Chen et al. [2013, 2016], Kveton et al. [2015], Wang and Chen [2017], Perrault et al. [2019b, 2020a]), particularly due to its wide applications in network routing, online advertising, recommender system, influence marketing, etc.

In CMAB, the whole joint distribution of the vector of outcomes $\mathbf{X}$ matters, contrary to standard MAB where only the marginals are sufficient to characterize a problem instance. For example, the following two extreme problem instances are distinct within the CMAB framework:

- (i) Each $\mathbb{P}_{X_i}$ is sub-Gaussian and the arm distributions are mutually independent, i.e., $\mathbb{P}_{\mathbf{X}} = \otimes_{i\in[n]}\mathbb{P}_{X_i}$.
- (ii) Each $\mathbb{P}_{X_i}$ is sub-Gaussian but the stochastic dependencies between the arm distributions are "worst case": the performance metric is the supremum of the regret over all possible dependencies between the marginals.

Those two settings are indeed different as two different lower bounds on the asymptotic (in T) regret can be derived. In particular, the regret scales as $\Omega(n\log(T)/\Delta)$ for the setting (i), and as $\Omega(mn\log(T)/\Delta)$ for (ii), where Δ is the minimum gap in the expected reward between an optimal super arm and any non-optimal super arm, and where $m \triangleq \max_{A\in\mathcal{A}}|A|$.

Many CMAB policies are based on the *Upper Confidence Bound* (UCB) approach, extending the classical UCB policy [Auer et al., 2002] from MAB to CMAB. This type of approach uses an optimistic estimate $\boldsymbol{\mu}_t$ of $\boldsymbol{\mu}^*$ (i.e., for which the reward function is overestimated), lying in a well-chosen confidence region. For setting (ii), there exist UCB-style policies that match the lower bound mentioned above. An example of such policy is *Combinatorial Upper Confidence Bound* (CUCB) [Chen et al., 2013, Kveton et al., 2015], that uses a Cartesian product of the individual confidence intervals of each arm as a confidence region. For setting (i), Combes et al. [2015] provided the UCB-style policy *Efficient Sampling for Combinatorial Bandit* (ESCB), that uses the assumption of mutual independence between arm distributions in order to build a tighter ellipsoidal confidence region around the empirical mean, which helps to better restrict the exploration. Degenne and Perchet [2016b] gave the following generalization of setting (i):

- (iii) The joint probability $\mathbb{P}_{\mathbf{X}}$ is $\mathbf{C}$-sub-Gaussian, for a positive semi-definite matrix $\mathbf{C} \succeq 0$, i.e., $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal\mathbf{C}\boldsymbol{\lambda}/2}$, for all $\boldsymbol{\lambda} \in \mathbb{R}^n$.

In this case, they provided the policy OLS-UCB, leveraging this additional assumption and such that it essentially reduces to ESCB in the specific case of diagonal matrix $\mathbf{C}$ with a regret bound of $\mathcal{O}\big(\log^2(m)n\log(T)/\Delta\big)$ (so it matches the above lower bound up to a polylogarithmic factor in m). We refer the reader to Table 1 for an overview of the above regret (lower) bounds.

In some CMAB problems, the action space $\mathcal{A}$ and the reward function are simple enough for the existence of an exact *oracle* that takes as input a vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and outputs the solution of the combinatorial problem (associated to the mean vector $\boldsymbol{\mu}$), with a polynomial time complexity $\mathcal{O}(\text{poly}(n))$. Under this assumption (referred to as Assumption 1), CUCB, that plays the action $A_t = \text{Oracle}(\boldsymbol{\mu}_t)$ at round t, is efficient to implement, and has a $\mathcal{O}(\text{poly}(n))$ time complexity per round. In that case, the setting (ii) is therefore essentially solved. On the other hand, this is not true for the settings (i) and (iii), as ESCB needs to solve a difficult combinatorial problem in each round (NP-Hard in general [Atamtürk and Gómez, 2017]).

The inefficiency of ESCB triggered some attempts to implement an efficient version: Perrault et al. [2019a] proposed an efficient approximation method for implementing ESCB in the case the action space has a *matroid* structure: they prove a time complexity of $\mathcal{O}(\text{poly}(n))$ while keeping the same

[2] Henceforth, we typeset vectors in bold and indicate components with indices, i.e., $\mathbf{a} = (a_i)_{i\in[n]} \in \mathbb{R}^n$. We also let $\mathbf{e}_i$ be the i^{th} canonical unit vector of $\mathbb{R}^n$, and define the incidence vector of any subset $A \subset [n]$ as $\mathbf{e}_A \triangleq \sum_{i\in A}\mathbf{e}_i$. We denote by $\mathbf{a} \odot \mathbf{b} \triangleq (a_i b_i)$ the Hadamard product of two vectors $\mathbf{a}$ and $\mathbf{b}$.

Table 1: Factor in front of $n\log(T)/\Delta$ in the regret bound ($\mathcal{O}(\cdot)$ for upper bounds), computationally inefficient policies are printed with a subscript $*$, setting (iii) is for $\mathbf{C}$ diagonal, CLIP CTS-GAUSSIAN is for linear reward functions, and with only $\boldsymbol{\lambda} \in \mathbb{R}_+^n$ in (iii). Our results are printed in bold, see Theorem 1, Theorem 2, Theorem 3 related to CTS-BETA, CTS-GAUSSIAN, CLIP CTS-GAUSSIAN respectively.

	CUCB	ESCB$_*$	CTS-BETA	CTS-GAUSSIAN	CLIP CTS-GAUSSIAN	Lower bound
(i)	m	$\log^2(m)$	$\mathbf{\log^2(m)}$	$\mathbf{\log^2(m)}$	$\mathbf{\log^2(m)}$	$\Omega(1)$
(ii)	m	m	-	$\mathbf{\log^2(m)m}$	$\mathbf{m}$	$\Omega(m)$
(iii)	m	$\log^2(m)$	-	$\mathbf{\log^2(m)}$	$\mathbf{\log^2(m)}$	$\Omega(1)$

regret rate. However, this improvement is mitigated by the fact that CUCB reaches the optimal regret rate $\mathcal{O}(n\log(T)/\Delta)$ for the special case of matroid semi-bandits [Anantharam et al., 1987, Kveton et al., 2014, Talebi and Proutiere, 2016]. Recently, Cuvelier et al. [2020] provided another approach for approximating ESCB for a wide variety of action spaces, including the matching bandit setting [Gai et al., 2010] and the online shortest path problem [Liu and Zhao, 2012], where CUCB is not known to be better than ESCB. However, their policies are still computationally expensive when T is large, since the time complexity at round t is of order $\mathcal{O}(t \cdot \text{poly}(n))$.

Another line of research is to find an efficient alternative to ESCB. One of the most promising candidate is *Thompson Sampling* (TS). Although introduced much earlier by Thompson [1933], the theoretical analysis of TS for frequentist MAB is quite recent: Kaufmann et al. [2012], Agrawal and Goyal [2012] gave a regret bound matching the UCB policy theoretically. Moreover, TS often performs better than UCB in practice, making TS an attractive policy for further investigations. For CMAB, TS extends to *Combinatorial Thompson Sampling* (CTS). In CTS, the unknown mean $\boldsymbol{\mu}^*$ is associated with a belief (a prior distribution) updated to a posterior with the Bayes'rule, each time a feedback is received. In order to choose an action at round t, CTS draws a sample $\boldsymbol{\theta}_t$ from the current belief, and plays the action given by $\text{Oracle}(\boldsymbol{\theta}_t)$. CTS is attractive also because its time complexity is $\mathcal{O}(\text{poly}(n))$ under Assumption 1. Recently, for the setting (i) with bounded outcomes, Wang and Chen [2018] proposed an analysis of CTS-BETA, which is CTS where the prior distribution is chosen to be a product of n Beta distributions. They proved two regret upper bounds depending on the class of reward functions:

$$\mathcal{O}\left(\frac{n\sqrt{m}\log(T)}{\Delta}\right) \text{ in the linear case and } \mathcal{O}\left(\frac{nm\log(T)}{\Delta}\right) \text{ in the general case.} \tag{1}$$

Although the aforementioned upper bound in the linear reward case outperforms the one of CUCB, it doesn't match the one of ESCB. To summarize, and despite many efforts, the existence of a policy that is both optimal (up to a polylogarithmic factor in m) and efficient in the setting (i) or (iii) is still an open problem, which we tackle in this paper.

Further related work We refer the reader to Wang and Chen [2018] for further related work on TS for combinatorial bandits, and particularly for Gopalan et al. [2014], that provided a frequentist high-probability regret bounds for TS with a general action space and a general feedback model — Komiyama et al. [2015], that investigated TS for the m-sets action space — Wen et al. [2015], that studied TS for contextual CMAB problems, using the Bayesian regret metric (see also Russo and Van Roy [2016]).

1.1 Contributions

We first improve the result of Wang and Chen [2018] by providing the regret upper bound $\mathcal{O}\big(\log^2(m)n\log(T)/\Delta\big)$ for CTS-BETA in the setting (i) with bounded outcomes. This bound is valid even for non linear reward functions. Our main contribution is a regret bound for the setting (iii). We propose an efficient policy called CTS-GAUSSIAN, that is CTS where the prior distribution is chosen to be a multivariate Gaussian. An analysis of CTS-GAUSSIAN allows us to obtain a regret bound reducing to $\mathcal{O}\big(\log^2(m)n\log(T)/\Delta\big)$ for a diagonal sub-Gaussian matrix. When the reward function is linear, we generalize the setting (iii) assuming only $\boldsymbol{\lambda} \in \mathbb{R}_+^n$. This allows us to get rid of negative correlations between the outcomes (as in Perrault et al. [2020b]), and focus on positive correlations. We propose in this setting the policy CLIP CTS-GAUSSIAN, where the score is truncated from below with the empirical mean, and from above with the UCB. Truncations from above are not

Algorithm 1 CTS-BETA

Initialization: For each arm i, let $a_i = b_i = 1$.
For all $t \geq 1$:
  Draw $\boldsymbol{\theta}_t \sim \otimes_{i\in[n]}\text{Beta}(a_i, b_i)$, and play $A_t = \text{Oracle}(\boldsymbol{\theta}_t)$.
  Get the observation $\mathbf{X}_t \odot \mathbf{e}_{A_t}$, and draw $\mathbf{Y}_t \sim \otimes_{i\in A_t}\text{Bernoulli}(X_{i,t})$.
  For all $i \in A_t$ update $a_i \leftarrow a_i + Y_{i,t}$ and $b_i \leftarrow b_i + 1 - Y_{i,t}$.

necessary, but can limit optimism, especially when positive correlations are significant. We obtain an improved regret bound for CLIP CTS-GAUSSIAN, where negative correlations no longer appear in the regret bound and where, in setting (ii), the extra $\log^2(m)$ factor present in the regret bound of CTS-GAUSSIAN disappears. All these results are summarized and compared to other state-of-the-art policies in Table 1.

2 Model

CMAB is formally introduced as follows. Consider a random process $(\mathbf{X}_t) \overset{iid}{\sim} \mathbb{P}_{\mathbf{X}}$, where $\mathbb{P}_{\mathbf{X}}$ is a distribution — unknown to the agent — of random vectors in $\mathbb{R}^n$, with unknown mean $\boldsymbol{\mu}^*$. At each round $t \in [T]$, the agent chooses a super arm (or action) $A_t \in \mathcal{A} \subset \mathcal{P}([n])$ based on the history of observations $\mathcal{H}_t \triangleq \sigma\big(\mathbf{X}_1 \odot \mathbf{e}_{A_1}, \dots, \mathbf{X}_{t-1} \odot \mathbf{e}_{A_{t-1}}\big)$ and a possible extra source of randomness (we denote by $\mathcal{F}_t$ the filtration containing $\mathcal{H}_t$ and the extra randomness of round t — in particular, $A_t \in \mathcal{F}_t$). The feedback received is then $\mathbf{X}_t \odot \mathbf{e}_{A_t}$ and the associated expected reward of the agent at that stage is $r(A_t, \boldsymbol{\mu}^*)$, for some known function r. The objective of the agent is to minimize the regret, defined for a policy π as

$$\forall T \geq 1, \quad R_T(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t\right],$$

where $\Delta_t \triangleq \Delta(A_t) \triangleq r(A^*, \boldsymbol{\mu}^*) - r(A_t, \boldsymbol{\mu}^*)$ with $A^* \in \arg\max_{A'\in\mathcal{A}} r(A', \boldsymbol{\mu}^*)$. As stated in the introduction, we will assume the following:

Assumption 1. *The agent has access to an oracle with a time complexity $\mathcal{O}(\text{poly}(n))$ such that for any mean vector $\boldsymbol{\mu}$,* $\text{Oracle}(\boldsymbol{\mu}) \in \arg\max_{A\in\mathcal{A}} r(A, \boldsymbol{\mu})$.

Similar to Chen et al. [2016], we assume that the function r satisfies the following smoothness property.

Assumption 2. *There exists a constant B, such that for every super arm $A \in \mathcal{A}$ and every pair of mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, $|r(A, \boldsymbol{\mu}) - r(A, \boldsymbol{\mu}')| \leq B\|\mathbf{e}_A \odot (\boldsymbol{\mu} - \boldsymbol{\mu}')\|_1$.*

For an arm $i \in [n]$, we define the number of time i has been chosen at the beginning of round t as $N_{i,t-1} \triangleq \sum_{t'\in[t-1]} \mathbb{I}\{i \in A_{t'}\}$. We also define the following quantities, that will be useful in the expression of an upper bound on the regret:

$m^* \triangleq \min_{A\in\arg\max_{A'\in\mathcal{A}} \mathbf{e}_{A'}^\intercal \boldsymbol{\mu}^*} |A|$ is the minimum size of an optimal action,
$\Delta_{i,\min} \triangleq \min_{A\in\mathcal{A},\ \Delta(A)>0,\ i\in A} \Delta(A)$, is the minimal gap of an action containing $i \in [n]$,
$\Delta_{\min} \triangleq \min_{i\in[n]} \Delta_{i,\min}$, is the minimal arm-gap and
$\Delta_{\max} \triangleq \max_{A\in\mathcal{A}} \Delta(A)$ is the maximal gap.

3 Regret bound for CTS-BETA in setting (i)

In this section, we consider the following assumption on top of the CMAB setting from section 2.

Assumption 3. *The outcomes X_i are bounded (in $[0, 1]$, w.l.o.g.), and are mutually independent (we are thus in a special case of (i)).*

For this problem, we consider CTS-BETA in Algorithm 1, which is described as follows. The prior is set to be a product of n beta distributions (being thus uniform over $[0, 1]$ initially). Notice, this prior is conjugate to a product of Bernoulli distributions. After the agent get an observation $X_{i,t}$,

Algorithm 2 CTS-GAUSSIAN

Input: The vector $\mathbf{D}$, and a parameter $\beta > 1$.
Initialization: Play each arm once (if the agent knows that $\boldsymbol{\mu}^* \in [a,b]^n$, this might be skipped)
For every subsequent round t:
Draw $\boldsymbol{\theta}_t \sim \otimes_{i\in[n]}\mathcal{N}\big(\overline{\mu}_{i,t-1}, N_{i,t-1}^{-1}\beta D_i\big)$ ($\theta_{i,t} \sim \mathcal{U}[a,b]$ if $N_{i,t-1}=0$).
Play $A_t = \text{Oracle}(\boldsymbol{\theta}_t)$.
Get the observation $\mathbf{X}_t \odot \mathbf{e}_{A_t}$, and update $\overline{\boldsymbol{\mu}}_{t-1}$ and counters accordingly.

it first binarizes it by sampling $Y_{i,t} \sim \text{Bernoulli}(X_{i,t})$ (the regret of the problem defined by the observations $Y_{i,t}$ is the same because $\mathbb{E}[Y_{i,t}] = \mu_i^*$). Then the prior is updated using Bayes' rule with each sample $Y_{i,t}$. When choosing a super arm at round t, the agent draws $\boldsymbol{\theta}_t$ from the beta belief, and then plugged it into the oracle, which outputs the super arm A_t to play.

The main result of this section is Theorem 1, that improves the regret bound of Wang and Chen [2018] for CTS-BETA.

Theorem 1. *The policy π described in Algorithm 1 has regret $R_T(\pi)$ of order*

$$\mathcal{O}\left(\sum_{i\in[n]} \frac{B^2 \log^2(m)\log(T)}{\Delta_{i,\min}}\right).$$

The proof of Theorem 1, as well as the complete non-asymptotic upper-bound is postponed to Appendix A. Our analysis incorporates two novelties that we detail in the two following paragraphs.

An improved leading term (cf. Step 3 of the proof of Theorem 1 in Appendix A) We define the *empirical average* of each arm $i \in [n]$ at the beginning of round t as $\overline{\mu}_{i,t-1} \triangleq \sum_{t'\in[t-1]} \frac{\mathbb{I}\{i\in A_{t'}\}Y_{i,t'}}{N_{i,t-1}}$. Notice that this empirical average definition differs from the one that is classically used in CMAB, since samples $Y_{i,t'}$ are used rather than $X_{i,t'}$. The improved dependence in m in the leading term of Theorem 1 (compared to (1)) is a consequence of two ingredients. The first is the following concentration inequality (see Appendix A, Lemma 2), which improves that of Wang and Chen [2018] by extending it to the case of non-linear reward. Indeed, we rather control the ℓ_1 norm in this case, instead of the ℓ_∞-norm, which leads to a tighter bound.

$$\mathbb{P}\left[\left\|\mathbf{e}_{A_t}\odot\left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1 \geq \sqrt{\frac{1}{2}\log(|\mathcal{A}|2^m T)\sum_{i\in A_t}\frac{1}{N_{i,t-1}}}\,\middle|\,\mathcal{H}_t\right] \leq 1/T. \tag{2}$$

The second ingredient is a more careful handling of the square-root term in the above probability, based on a method similar to the one in Degenne and Perchet [2016b].

T-independent term (cf. Step 4 of the proof of Theorem 1 in Appendix A) Similarly to Wang and Chen [2018], our regret bound also contains an exponential term that is constant in T. Note, however that the term of Wang and Chen [2018] is of order $\mathcal{O}(\varepsilon^{-2m^*-2})$, whereas ours is of order $\mathcal{O}\big(\varepsilon^{-4m^*-2}\big)$, where $\varepsilon \in (0,1)$ is of order $\Delta_{\min}/(m^*)^2$. This discrepancy is due to the correction of a minor negligence inaccuracy in their Lemma 7, where they assume, at the end of the proof, that one could decorrelate the counters from the outcomes received. We manage to circumvent this issue by doing a careful union bound over the counters. It is this union bound that brings a larger dependence in this constant term. An additional discussion is deferred to the end of Appendix A.

4 Regret bound for CTS-GAUSIAN in setting (*iii*)

In this section, we consider the setting from section 2, with a more general sub-Gaussian family for $\mathbf{X} \in \mathbb{R}^n$. More precisely, we make the following similar assumption as in Degenne and Perchet [2016b]. Proposition 1 gives two examples included in this assumption (see Appendix B for a proof).

Assumption 4. *There exists a vector $\mathbf{D} \triangleq (D_1,\dots,D_n) \in \mathbb{R}_+^n$ known to the agent such that*

$$\forall A \in \mathcal{A},\ \forall \boldsymbol{\lambda}\in\mathbb{R}^n\ s.t.\ \boldsymbol{\lambda} = \boldsymbol{\lambda}\odot\mathbf{e}_A,\quad \mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal\mathbf{D}\odot\boldsymbol{\lambda}/2}.$$

Motivation for sub-Gaussian outcomes In the same way as boundedness generalizes to sub-Gaussianity in 1d, we have that if $\mathbf{X}$ is a.s. in a compact $\mathcal{K}$, it is $\mathbf{C}$-sub-Gaussian, with $\mathbf{C}$ built from the John's ellipsoid of $\mathcal{K}$. In this case, D_i is computed with a linear maximization over $\mathcal{A}$. In particular, $\mathcal{K} = B_{\ell_\infty}(0,1)$ gives $D_i = m$, and $\mathcal{K} = B_{\ell_2}(0,1)$ gives $D_i = 1$. We can also use other structures on the outcomes to have D_i, such as negative dependence (as we will see in our shortest path experiments, in section 5).

Proposition 1. *Assumption 4 encompasses the κ_i^2-sub Gaussian outcomes with worst case dependencies between the arm distributions, taking $D_i = \kappa_i^2 m$. It also captures $\mathbf{C}$-sub-Gaussian outcomes with a known sub-Gaussian matrix $\mathbf{C}$ (setting (iii)), taking $D_i = \max_{A\in\mathcal{A},\, i\in A} \sum_{j\in A} |C_{ij}|$.*

For the above setting, we provide CTS-GAUSSIAN in Algorithm 2, where we define the empirical mean of arm i at round $t \geq 1$ as $\overline{\mu}_{i,t-1} \triangleq \sum_{t'\in[t-1]} \frac{\mathbb{I}\{i\in A_{t'}\} X_{i,t'}}{N_{i,t-1}}$. This algorithm is comparable to Algorithm 1 but considers a Gaussian prior for each arm. Notice, the Gaussian family is *self-conjugate*, so except in the Gaussian-outcomes case, we do not rely on exact conjugated prior here. Although this is not surprising — since it is known that TS can work without exact conjugate prior with respect to the outcomes — obtaining an upper bound on the regret of the policy CTS-GAUSSIAN is non-trivial and constitutes our main contribution. We state our main result in Theorem 2.

Theorem 2. *The policy π described in Algorithm 2 has regret $R_T(\pi)$ of order*

$$\mathcal{O}\left(\sum_{i\in[n]} \frac{B^2 D_i \log^2(m)\log(T)}{\Delta_{i,\min}}\right).$$

The proof of Theorem 2, as well as the complete non-asymptotic upper-bound is postponed to Appendix C. Nonetheless, in the following paragraphs, we provide some insights and highlight the novelty of our analysis.

Main proof challenges In the setting of the previous section, the outcomes are independent in $[0,1]$ and an important step in Algorithm 1 was to transform the outcomes into binary variables in order to be consistent with the posterior. Here, outcomes are no longer independent. In addition to that, we cannot transform the outcomes into Gaussian variables in the same way as in Algorithm 1. These two points are the main technical challenges to address in our analysis.

Stochastic dominance Before providing details on how we deal with the above challenges, first recall that the standard analysis (in the case of a factorized prior, that we have here[3]) consists in bounding the expected number of rounds needed for the sample $\boldsymbol{\theta}_t$ to be close to the true mean $\boldsymbol{\mu}^*$ on a certain set $Z \subset A^*$, i.e., for the event $\{\|(\boldsymbol{\mu}^* - \boldsymbol{\theta}_t)\odot \mathbf{e}_Z\|_\infty > \varepsilon\}$ to happen. We let $\mathfrak{T}_t(Z)$ denote the complementary event. As for the proof of Theorem 1, we can condition on the history to rewrite this expected number of rounds and then upper bound it as

$$\mathbb{E}\left[\sum_{t\geq 1}(t-1)\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]\prod_{j=1}^{t-1}\mathbb{P}[\mathfrak{T}_j(Z)|\mathcal{H}_j]\right]$$
$$\leq \mathbb{E}\left[\sup_{t\geq 1}\frac{1}{\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]}\right] - 1 \leq \sum_{Z'\subset Z,\ Z'\neq\emptyset}\mathbb{E}\left[\sup_{t\geq 1}\prod_{i\in Z'}\left(\frac{1}{\mathbb{P}[|\theta_{i,t}-\mu_i^*|\leq\varepsilon|\mathcal{H}_t]} - 1\right)\right].$$

Now, using the fact that the conditional distribution of $\theta_{i,t} - \overline{\mu}_{i,t-1}$ is symmetric and depends only on the counter $N_{i,t-1}$, we obtain that the probability $\mathbb{P}[|\theta_{i,t} - \mu_i^*| \leq \varepsilon|\mathcal{H}_t]$ is a monotonic function of the deviation $\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|$. Let us emphasize that this property of the Gaussian prior used is crucial and that it is not obvious to transfer the same technique to a beta prior. To sum up, we have to control a term of the form $\mathbb{E}\left[\sup_{t\geq 1}\prod_{i\in Z'} g_i\left(\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|\right)\right]$, where g_i are non-negative increasing functions. Our approach is to prove that $\left(\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|\right)_i$ is *weakly stochastically dominated* by $\left(\sqrt{\frac{\beta D_i}{N_{i,t-1}}}|\eta_i|\right)_i$, where $\boldsymbol{\eta} \sim \otimes_i \mathcal{N}(0,1)$, which is the same vector but where the empirical mean is

[3] In practice, for $\mathbf{C}$-sub Gaussian outcomes, the choice $\mathcal{N}\left(\overline{\boldsymbol{\mu}}_{t-1}, \left(C_{ij}N_{ij,t-1}N_{i,t-1}^{-1}N_{j,t-1}^{-1}\right)_{ij}\right)$ for the prior where $N_{ij,t-1} \triangleq \sum_{t'\in[t-1]}\mathbb{I}\{i\in A_{t'}\}\mathbb{I}\{j\in A_{t'}\}$ may be preferred.

built with independent Gaussian outcomes instead. Notice, independence is crucial to be able to factorize the expectation $\mathbb{E}\big[\prod_{i\in Z'} g_i\big]$, in the same way as in the proof of Theorem 1. We recall two equivalent definitions of $\mathbf{U}$ is weakly stochastically dominated by $\mathbf{V}$, see Shaked and Shanthikumar [2007] for more details and properties of dominances,

- For all non-negative, non-increasing functions f_i, it holds $\mathbb{E}[\prod_i f_i(U_i)] \leq \mathbb{E}[\prod_i f_i(V_i)]$.
- For any vector $\mathbf{x}$, it holds $\mathbb{P}[\mathbf{U} \geq \mathbf{x}] \leq \mathbb{P}[\mathbf{V} \geq \mathbf{x}]$.

The first point applied to g_i's (and up to the supremum over t) is a simple way to obtain the aforementioned wanted control. Thus, it's enough to prove the second point, which is a consequence of the sub-Gaussianity of outcomes given by Assumption 4 and some concentration inequality. Finally, we circumvent the supremum over $t \geq 1$ issue thanks to Doob's optional sampling theorem for non-negative super-martingales (see Durrett [2019], Theorem 5.7.6).

Importance of using a factorized prior in our analysis Note that in Algorithm 2, the samples $\theta_{i,t}$ are independent, while the outcomes are not necessarily independent. This independence is in fact crucial in order to be able to start the analysis in the same way as in the proof of Theorem 1 (recall that Algorithm 1 also uses a factorized prior). More precisely, a factorized prior allows us to link the filtered regret against the event $\mathfrak{S}_t(Z) \wedge \mathfrak{T}_t(Z)$ to the expected number of rounds needed for $\neg\mathfrak{T}_t(Z)$ to occur (see (3) in Step 4 of the proof of Theorem 1 in Appendix A for a definition of $\mathfrak{S}_t(Z)$). Indeed, without the factorized prior, the two events $\mathfrak{S}_t(Z), \mathfrak{T}_t(Z)$ would no longer be independent conditionally to the history, and the term $1/\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]$ obtained in the previous paragraph would then be replaced by $1/\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathfrak{S}_t(Z), \mathcal{H}_t]$, which is much more difficult to deal with. To the best of our knowledge, it is unknown how to get the desired bound when $\mathfrak{S}_t(Z)$ and $\mathfrak{T}_t(Z)$ are not independent conditionally to the history.

4.1 CLIP CTS-GAUSSIAN for the linear reward case

In this subsection, we make the following assumptions on top of Section 2.

Assumption 5. *The reward function is linear, defined as* $r(A, \boldsymbol{\mu}) \triangleq \mathbf{e}_A^\intercal \boldsymbol{\mu}$.

Assumption 6. *The agent knows a matrix* $\boldsymbol{\Gamma} \succeq 0$ *s.t.* $\forall \boldsymbol{\lambda} \in \mathbb{R}_+^n$, $\mathbb{E}\big[e^{\boldsymbol{\lambda}(\mathbf{X}-\boldsymbol{\mu}^*)}\big] \leq e^{\boldsymbol{\lambda}^\intercal \boldsymbol{\Gamma} \boldsymbol{\lambda}/2}$.

Notice that Assumption 6 slightly generalises the setting from Degenne and Perchet [2016b]. Requiring $\boldsymbol{\lambda} \in \mathbb{R}_+^n$ allows us to take $D_i = \max_{A\in\mathcal{A},\, i\in A} \sum_{j\in A}(0 \vee \Gamma_{ij})$, so that negative correlations are no longer harmful. D_i can still be too large (and thus $\boldsymbol{\theta}_t$ might be over-sampled), so we cap $\boldsymbol{\theta}_t$ with the score $\boldsymbol{\mu}_t$ used by CUCB. The resulting policy is CLIP CTS-GAUSSIAN, where the score $\boldsymbol{\theta}_t$ is replaced by $\overline{\boldsymbol{\mu}}_{t-1} \vee \boldsymbol{\theta}_t \wedge \boldsymbol{\mu}_t$ before we plug it into Oracle, where $\mu_{i,t} = \overline{\mu}_{i,t-1} + \sqrt{\Gamma_{ii}\frac{2(\log(t)+4\log\log(t))}{N_{i,t-1}}}$. CLIP CTS-GAUSSIAN enjoys the following regret bound.

Theorem 3. *The policy* CLIP CTS-GAUSSIAN *has regret of order*

$$\mathcal{O}\left(\sum_{i\in[n]} \frac{\big(D_i \log^2(m) \wedge m\Gamma_{ii}\big)\log(T)}{\Delta_{i,\min}}\right).$$

Not only D_i is improved through the above relaxation, but also, the leading term is never worse than the one of CUCB. The proof and the complete non-asymptotic upper-bound is delayed to Appendix D. We note that we rely heavily on reward linearity to analyse this clip version, not only using monotony to restrict the controls to the $\mathbb{R}_+^n$ directions (and thus to cap from bellow the sample by the empirical mean), but also using the oracle's invariance property $\text{Oracle}(\boldsymbol{\mu}) = \text{Oracle}\big(\boldsymbol{\mu} + \boldsymbol{\delta} \odot \mathbf{e}_{\text{Oracle}(\boldsymbol{\mu})}\big)$, with $\boldsymbol{\delta} \geq 0$, to cap the sample from above by the UCB.

Comparison with the OLS-UCB analysis of Degenne and Perchet [2016b] The leading term in the regret bound given from Theorem 3 is comparable to the one for OLS-UCB from Degenne and Perchet [2016b]. Indeed, we recall that they obtained a factor of order $\Gamma_{ii}\big((1-\gamma)\log^2(m) + \gamma m\big)$, with $\gamma \triangleq \max_{A\in\mathcal{A}} \max_{(i,j)\in A^2, i\neq j}(0 \vee \Gamma_{ij})/\sqrt{\Gamma_{ii}\Gamma_{jj}}$, where we have $\big(D_i \log^2(m) \wedge m\Gamma_{ii}\big)$. When $\gamma \in \{0, 1\}$ (this is the case when we are in the settings (i) and (ii) respectively), these two terms coincide. When $\gamma \in (0, 1)$, they are incomparable in general. We can still see that our variance term D_i is always lower than their $\Gamma_{ii}((1-\gamma) + \gamma m)$, i.e., that our bound rate is lower than $\log^2(m)$ times theirs.

5 Experiments

Before describing the experiments carried out, notice that in the CTS-GAUSSIAN policies, $\beta > 1$ is an artefact of the analysis and can in practice be taken equal to 1. This is what we did in our experiments.

The shortest path problem We compare our CTS policies to CUCB and CUCB-KL, for the shortest path problem on the road chesapeake network [Rossi and Ahmed, 2015]. This network contains 39 nodes and $n = 170$ edges. $\mathcal{A}$ is the set of paths from an origin to a destination in the network. We choose a linear reward, so that an efficient Oracle exists for this problem. We choose $\boldsymbol{\mu}^*$ uniformly in $[-1, 0]^n$ and then normalize its sum so that $\sum_i \mu_i^* = -s$, where s is unknown to the agent. The parameter s stands for the global network traffic (e.g., the total number of vehicles in the network). We run two experiments, one with $-\mathbf{X} \sim \otimes_i \text{Bernoulli}(-\mu_i^*)$ and another with $-\mathbf{X} \sim \otimes_i \text{Bernoulli}(-\mu_i^*)$ conditionally on $\sum_i X_i = -s$. They are presented in Figure 1. Since the outcomes are not mutually independent in this last experiment, we use (CLIP) CTS-GAUSSIAN rather than CTS-BETA, where we take $D_i = 1/4$, using that for any $\boldsymbol{\lambda} \in \mathbb{R}_+^n$, $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal \mathbf{X}}\right] \leq \prod_{i\in[n]} \mathbb{E}\left[e^{\lambda_i X_i}\right]$ (see e.g., Borcea et al. [2009], corollary 4.18). It is clear from the experiments that CTS policies outperform both CUCB and CUCB-KL. In the second experiment, we see that CLIP CTS-GAUSSIAN and CTS-GAUSSIAN are very similar — which is not surprising because D_i is not large here (unlike in the next experiment) — and that for a small s, CUCB-KL becomes competitive, since the kl is much larger than the quadratic divergence in that case.

Comparison to ESCB for the matching problem We consider here a comparison between (CLIP) CTS-GAUSSIAN, CUCB and ESCB (we refer the reader to Wang and Chen [2018] for a comparison between CTS-BETA and ESCB). Since ESCB is computationally intractable, we limit ourselves to a toy matching problem on the complete bipartite graphs $K_{4,4}$, with $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^*, (c\mathbb{I}\{i \neq j\} + \mathbb{I}\{i = j\})_{ij})$, where this covariance is known to the agent. Our results are shown in Figure 2, where we observe that CLIP CTS-GAUSSIAN (resp. ESCB) is slightly better for c small (resp. large), thus reaching the best of both worlds. This is because a large c forces CLIP CTS-GAUSSIAN to oversample (as evidenced by CTS-GAUSSIAN whose performance is even worse than CUCB for $c = 1$). We also recorded the computation time for larger instances (see Table 2), and observe the efficiency of CUCB and CLIP CTS-GAUSSIAN compared to ESCB.

Correlated vs independent prior in practice We briefly discussed the use of a correlated prior in footnote 3, with covariance $\left(C_{ij} N_{ij,t-1} N_{i,t-1}^{-1} N_{j,t-1}^{-1}\right)_{ij}$, mentioning that the policy would perform better than using an independent prior. We ran additional empirical comparisons to assess this, plotting the results in Figure 3 where we also compared with a common prior policy approach [Agrawal et al., 2017], i.e., with covariance $\left(N_{i,t-1}^{-1/2} N_{j,t-1}^{-1/2}\right)_{ij}$.[4] As expected, the correlated prior policy is better than the independent one (when outcomes are correlated). This motivates the theoretical study of such policy for future work. The common prior approach is comparable to the correlated prior one on the matching problem, but it is outperformed in the worst-case scenario of a separate action space $\mathcal{A} = \left\{\{km + 1, \dots, (k + 1)m\} \mid k \in \left\{0, \dots, \frac{n}{m} - 1\right\}\right\}$ with independent outcomes. This is because such problem reduces to a classical MAB problem with a covariance scaled up by a factor m, whereas the common prior approach has a variance scaled up by a factor m^2.

Table 2: Computation time per round (ms), with $c = 0.3$, $T = 100$, averaged over 5 simulations.

	$K_{3,3}$	$K_{4,4}$	$K_{5,5}$	$K_{6,6}$	$K_{7,7}$	$K_{8,8}$
CUCB	0.39	0.64	1.23	1.65	2.45	3.88
CLIP CTS-GAUSSIAN	0.50	0.80	1.75	1.79	3.30	5.42
ESCB	0.45	1.93	10.3	75.6	541	4694

[4] We also tried the policy (without displaying the results, for the sake of clarity) with covariance $\left(C_{ij} N_{i,t-1}^{-1/2} N_{j,t-1}^{-1/2}\right)_{ij}$, and observed about the same performance as the correlated prior approach.

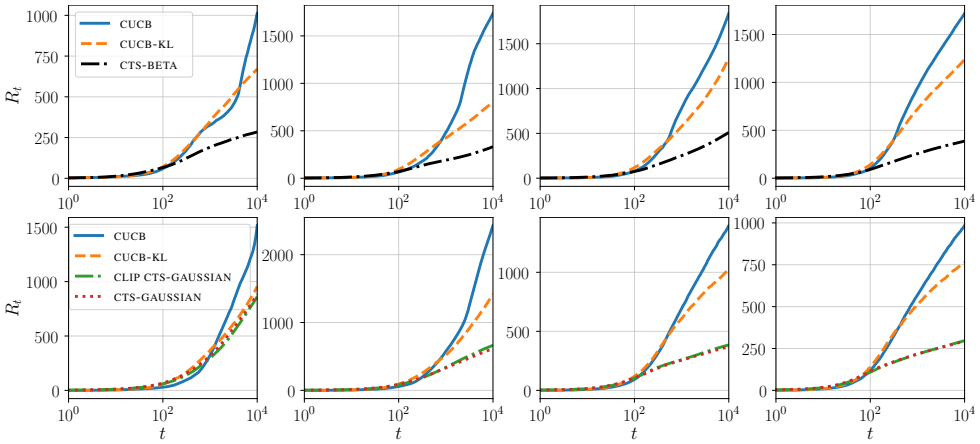

Figure 1: Cumulative regret (averaged over 50 simulations) for the shortest path problem. **Top:** with mutually independent outcomes, taking the opposite sum of means being $s = 70, 90, 110, 130$ respectively. **Bottom:** with correlated outcomes, taking the opposite sum of outcomes being $s = 70, 90, 110, 130$ respectively.

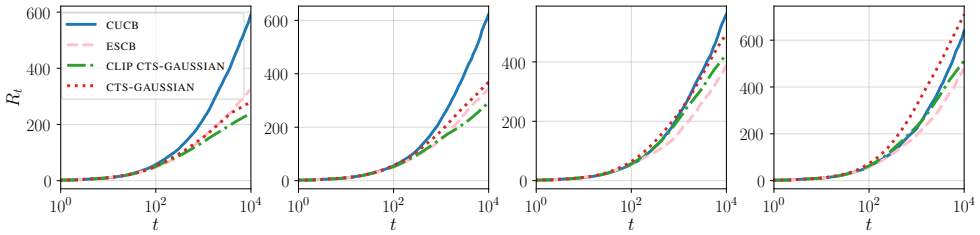

Figure 2: Cumulative regret (averaged over 50 simulations) for the matching problem with Gaussian outcomes, taking $c = -1/n, 0.2, 0.5, 1$ respectively.

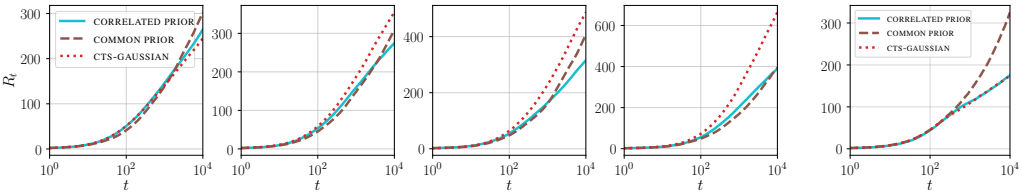

Figure 3: Comparison with correlated prior sampling and common prior sampling (averaged over 50 simulations). **The first 4:** for the $K_{4,4}$ matching problem, with Gaussian outcomes, taking $c = 0, 0.2, 0.5, 1$. **The last:** for $\mathcal{A} = \big\{\{km+1, \ldots, (k+1)m\} \mid k \in \big\{0, \ldots, \frac{n}{m} - 1\big\}\big\}$, $c = 0$.

6 Conclusion and future work

In this paper, we have provided the first efficient policies having an optimal regret bound for a wide spectrum of problems instances for CMAB with semi-bandit feedback. Our approach also answers the question of finding an analysis for CTS under correlated arm distributions. There are several possible extensions that could be considered as future work. For example, it would be interesting to have an analysis of CTS with a *correlated* (Gaussian) prior. Indeed, apart from the empirical gain, this would open up the possibility of estimating the covariance matrix and using it in the prior distribution. Further relevant results would be an analysis of CTS-BETA without the mutual independence of outcomes, or also an improved concentration bound for a sum of independent betas, relying on the kl rather than using sub-Gaussianity. This latter result would thus show that CTS-BETA dominates CUCB-KL, which is empirically observed.

Broader Impact

This work does not present any foreseeable societal consequence.

Acknowledgments and Disclosure of Funding

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).

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

S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR, abs/1209.3352, http://arxiv.org/abs/1209.3352*, sep 2012. URL http://arxiv.org/abs/1209.3352.

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*, 2017.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

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

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

D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*, volume 38 of *Monographs on statistics and applied probability*. Chapman and Hall, 1985.

J. Borcea, P. Brändén, and T. Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2):521–567, 2009.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. In *Journal of Computer and System Sciences*, volume 78, pages 1404–1422, 2012.

S.-H. Chang, P. C. Cosman, and L. B. Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/chen13a.html.

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

R. Combes, M. S. T. M. Shahi, A. Proutiere, and Others. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015.

T. Cuvelier, R. Combes, and E. Gourdin. Statistically efficient, polynomial time algorithms for combinatorial semi bandits. *arXiv preprint arXiv:2002.07258*, 2020.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1587–1595, New York, New York, USA, 20–22 Jun 2016a. PMLR. URL http://proceedings.mlr.press/v48/degenne16.html.

R. Degenne and V. Perchet. Combinatorial semi-bandit with known covariance. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2972–2980. Curran Associates, Inc., 2016b. URL http://papers.nips.cc/paper/6137-combinatorial-semi-bandit-with-known-covariance.pdf.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, pages 1–9. IEEE, 2010.

Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *Transactions on Networking*, 20(5):1466–1478, 2012.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex bandit problems. In *International Conference on Machine Learning*, 2014.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

I. M. Jacobs and J. Wozencraft. Principles of communication engineering. 1965.

E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. *Algorithmic Learning Theory*, 2012.

J. Komiyama, J. Honda, and H. Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. jun 2015. URL http://arxiv.org/abs/1506.00779.

B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Uncertainty in Artificial Intelligence*, 2014.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, 2015.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

K. Liu and Q. Zhao. Adaptive shortest-path routing under unknown and stochastically varying link states. In *2012 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pages 232–237. IEEE, 2012.

O. Marchal, J. Arbel, et al. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22, 2017.

P. Perrault, V. Perchet, and M. Valko. Exploiting structure of uncertainty for efficient matroid semi-bandits. In K. Chaudhuri and R. 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.

P. Perrault, V. Perchet, and M. Valko. Finding the bandit in a graph: Sequential search-and-stop. In K. Chaudhuri and M. 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.

P. Perrault, J. Healey, Z. Wen, and M. Valko. Budgeted online influence maximization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6588–6599. 2020a.

P. Perrault, V. Perchet, and M. Valko. Covariance-adapting algorithm for semi-bandits with application to sparse rewards. In *Conference on Learning Theory*, 2020b.

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

R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015. URL http://networkrepository.com.

D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

M. S. Talebi and A. Proutiere. An Optimal Algorithm for Stochastic Matroid Bandit Optimization. In *The 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 548–556, 2016. ISBN 9781450342391.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

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

S. Wang and W. Chen. Thompson Sampling for Combinatorial Semi-Bandits. mar 2018. URL http://arxiv.org/abs/1803.04623.

Z. Wen, B. Kveton, and A. Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122, 2015.

A Proof of Theorem 1

We first restate the complete non-asymptotic upper-bound as follows.

Theorem. *The policy π described in Algorithm 1 has regret $R_T(\pi)$ bounded by*

$$16\log_2^2(16m)\sum_{i\in[n]}\frac{B^2\log(2^m|\mathcal{A}|T)}{\Delta_{i,\min}}+\Delta_{\max}(1+n)+\frac{nm^2\Delta_{\max}}{\left(\frac{\Delta_{\min}}{2B}-(m^{*2}+1)\varepsilon\right)^2}+\Delta_{\max}\frac{C}{\varepsilon^2}\left(\frac{C'}{\varepsilon^4}\right)^{m^*},$$

where C, C' are two universal constants, and $\varepsilon\in(0,1)$ is such that $\Delta_{\min}/(2B)-(m^{*2}+1)\varepsilon>0$.

A.1 Preliminary lemmas

In order to prove Theorem 1, we modify two lemmas from Wang and Chen [2018]: first, in their Lemma 3, we replace ε by $\Delta_{\min}/(2B)-(m^{*2}+1)\varepsilon>0$, which gives the following Lemma 1.

Lemma 1. *In Algorithm 1, for any arm i, we have*

$$\mathbb{E}\left[\left|t\in[T],\ i\in A_t,\ |A_t|\cdot\left|\overline{\mu}_{i,t-1}-\mu_i^*\right|>\frac{\Delta_{\min}}{2B}-(m^{*2}+1)\varepsilon\right|\right]\leq 1+\left(\frac{\Delta_{\min}}{2mB}-\frac{(m^{*2}+1)\varepsilon}{m}\right)^{-2}.$$

Then, we modify Lemma 4 from Wang and Chen [2018] as follows, leveraging on the mutual independence of $\theta_{1,t},\dots,\theta_{n,t}$ to get a tighter confidence region for the sample $\boldsymbol{\theta}_t$.

Lemma 2. *In Algorithm 1, for all round t, we have*

$$\mathbb{P}\left[\left\|\mathbf{e}_{A_t}\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\sqrt{\frac{1}{2}\log(|\mathcal{A}|2^mT)\sum_{i\in A_t}\frac{1}{N_{i,t-1}}}\,\middle|\,\mathcal{H}_t\right]\leq 1/T.$$

Proof. From [Marchal et al., 2017], the Beta random variable from $\theta_{i,t}$ is sub-Gaussian with variance $1/(4N_{i,t-1})$. Thus, defining the functions

$$\alpha_t(A)\triangleq\sqrt{\frac{1}{2}\log(|\mathcal{A}|2^mT)\sum_{i\in A}\frac{1}{N_{i,t-1}}},\quad\text{and}\quad\lambda_t(A)\triangleq\frac{4\alpha_t(A)}{\sum_{i\in A}1/N_{i,t-1}},$$

we have

$$\begin{aligned}
\mathbb{P}\left[\left\|\mathbf{e}_{A_t}\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\alpha_t(A_t)\middle|\mathcal{H}_t\right]&\leq\sum_{A\in\mathcal{A}}\mathbb{P}\left[\left\|\mathbf{e}_A\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\alpha_t(A)\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\mathbb{E}\left[e^{\lambda_t(A)\left\|\mathbf{e}_A\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\prod_{i\in A}\mathbb{E}\left[e^{\lambda_t(A)\left|\theta_{i,t}-\overline{\mu}_{i,t-1}\right|}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\prod_{i\in A}\mathbb{E}\left[e^{\lambda_t(A)\left(\theta_{i,t}-\overline{\mu}_{i,t-1}\right)}+e^{\lambda_t(A)\left(\overline{\mu}_{i,t-1}-\theta_{i,t}\right)}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}2^{|A|}e^{-\lambda_t(A)\alpha_t(A)}e^{\lambda_t(A)^2\sum_{i\in A}1/(8N_{i,t-1})}\leq 1/T.
\end{aligned}$$

□

A.2 Main proof

With the two lemmas from the previous subsection, we are ready to demonstrate Theorem 1. We consider the following events.

- $\mathfrak{Z}_t\triangleq\{\Delta_t>0\}$
- $\mathfrak{B}_t\triangleq\left\{\exists i\in A_t,\ |A_t|\cdot\left|\overline{\mu}_{i,t-1}-\mu_i^*\right|>\Delta_{\min}/(2B)-(m^{*2}+1)\varepsilon\right\}$

- $\mathfrak{C}_t \triangleq \left\{\|\mathbf{e}_{A_t} \odot (\boldsymbol{\theta}_t - \boldsymbol{\mu}^*)\|_1 > \Delta_t/B - \left(m^{*2}+1\right)\varepsilon\right\}$
- $\mathfrak{D}_t \triangleq \left\{\left\|\mathbf{e}_{A_t} \odot \left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1 \geq \sqrt{0.5 \cdot \log(|\mathcal{A}|2^m T)\sum_{i\in A_t} 1/N_{i,t-1}}\right\}$.

We break down our analysis into 4 steps. The main novelties are in the last two steps: Step 3 gives us the tighter dependence in m, and Step 4, that contains the main difficulties, gives the new exponential constant term.

Step 1: bound under $\mathfrak{Z}_t \wedge \mathfrak{B}_t$ By Lemma 1,

$$\sum_{t\in[T]} \mathbb{E}[\Delta_t \mathbb{I}\{\mathfrak{Z}_t \wedge \mathfrak{B}_t\}] \leq \Delta_{\max} \sum_{i\in[n]} \mathbb{E}\left[\left|t \in [T],\ i \in A_t,\ |A_t| \cdot \left|\overline{\mu}_{i,t-1} - \mu_i^*\right| > \Delta_{\min}/(2B) - (m^{*2}+1)\varepsilon\right|\right]$$
$$\leq n\Delta_{\max}\left(1 + \left(\frac{\Delta_{\min}}{2mB} - \frac{(m^{*2}+1)\varepsilon}{m}\right)^{-2}\right).$$

Step 2: bound under $\mathfrak{Z}_t \wedge \neg\mathfrak{B}_t \wedge \mathfrak{C}_t \wedge \mathfrak{D}_t$ By Lemma 2,

$$\sum_{t\in[T]} \mathbb{E}[\Delta(A_t)\mathbb{I}\{\mathfrak{Z}_t \wedge \neg\mathfrak{B}_t \wedge \mathfrak{C}_t \wedge \mathfrak{D}_t\}] \leq \Delta_{\max} \sum_{t\in[T]} \mathbb{E}[\mathbb{P}[\mathfrak{D}_t|\mathcal{H}_t]] \leq \Delta_{\max} \sum_{t\in[T]} 1/T = \Delta_{\max}.$$

Step 3: bound under $\mathfrak{Z}_t \wedge \neg\mathfrak{B}_t \wedge \mathfrak{C}_t \wedge \neg\mathfrak{D}_t$

$$\begin{aligned}
\Delta_t/B &\leq \|\mathbf{e}_{A_t} \odot (\boldsymbol{\theta}_t - \boldsymbol{\mu}^*)\|_1 + \left(m^{*2}+1\right)\varepsilon && \mathfrak{C}_t \\
&\leq \left\|\mathbf{e}_{A_t} \odot \left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1 + \left\|\mathbf{e}_{A_t} \odot \left(\overline{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*\right)\right\|_1 + \left(m^{*2}+1\right)\varepsilon \\
&\leq \left\|\mathbf{e}_{A_t} \odot \left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1 + \Delta_{\min}/(2B) - \left(m^{*2}+1\right)\varepsilon + \left(m^{*2}+1\right)\varepsilon && \neg\mathfrak{B}_t \\
&\leq \left\|\mathbf{e}_{A_t} \odot \left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1 + \Delta_t/(2B) && \mathfrak{Z}_t \\
&\leq \sqrt{\frac{1}{2}\log(|\mathcal{A}|2^m T)\sum_{i\in A_t}\frac{1}{N_{i,t-1}}} + \Delta_t/(2B). && \neg\mathfrak{D}_t
\end{aligned}$$

So we have that the following event holds

$$\mathfrak{A}_t \triangleq \left\{\Delta_t \leq B\sqrt{2\log(|\mathcal{A}|2^m T)\sum_{i\in A_t}\frac{1}{N_{i,t-1}}}\right\}.$$

We can thus apply Theorem 4 (see Appendix E) to get the bound

$$\sum_{t\in[T]} \mathbb{E}[\Delta_t \mathbb{I}\{\mathfrak{Z}_t, \neg\mathfrak{B}_t, \mathfrak{C}_t, \neg\mathfrak{D}_t\}] \leq \sum_{t\in[T]} \mathbb{E}[\Delta_t \mathbb{I}\{\mathfrak{A}_t\}]$$
$$\leq 32B^2 \log_2^2(4\sqrt{m}) \sum_{i\in[n]} \Delta_{i,\min}^{-1} 2\log(|\mathcal{A}|2^m T).$$

Step 4: bound under $\mathfrak{Z}_t \wedge \neg\mathfrak{C}_t$ We consider the following events for a subset $Z \subset [n]$

$$\mathfrak{R}(\boldsymbol{\theta}', Z) \triangleq \left\{Z \subset \text{Oracle}(\boldsymbol{\theta}'),\ \left\|\mathbf{e}_{\text{Oracle}(\boldsymbol{\theta}')} \odot \left(\boldsymbol{\theta}' - \boldsymbol{\mu}^*\right)\right\|_1 > \Delta\left(\text{Oracle}(\boldsymbol{\theta}')\right) - (k^{*2}+1)\varepsilon\right\}$$

$$\mathfrak{S}_t(Z) \triangleq \left\{\forall \boldsymbol{\theta}' \text{ s.t. } \left\|\left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_Z\right\|_\infty \leq \varepsilon,\ \mathfrak{R}(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \boldsymbol{\theta}_t \odot \mathbf{e}_{Z^c}, Z) \text{ holds}\right\} \tag{3}$$

$$\mathfrak{T}_t(Z) \triangleq \{\|(\boldsymbol{\mu}^* - \boldsymbol{\theta}_t) \odot \mathbf{e}_Z\|_\infty > \varepsilon\}.$$

We can state the three following lemmas. Note that Lemma 3 is exactly the Lemma 1 from Wang and Chen [2018]. The other two replace their Lemma 7.

Lemma 3. *In Algorithm 1, for all round t, we have*

$$\mathfrak{Z}_t, \neg\mathfrak{C}_t \Rightarrow \exists Z \subset A^*,\ Z \neq \emptyset \text{ s.t. the event } \mathfrak{S}_t(Z) \wedge \mathfrak{T}_t(Z) \text{ holds.}$$

Lemma 4. *Given $Z \subset A^*,\ Z \neq \emptyset$, let τ_q be the round at which $\mathfrak{S}_t(Z) \wedge \neg\mathfrak{T}_t(Z)$ occurs for the q-th time, and let $\tau_0 = 0$. Then, in Algorithm 1, we have*

$$\mathbb{E}\left[\sum_{t=\tau_q+1}^{\tau_{q+1}} \mathbb{I}\{\mathfrak{S}_t(Z), \mathfrak{T}_t(Z)\}\right] \leq \mathbb{E}\left[\sup_{\tau \geq \tau_q+1} \prod_{i \in Z} \frac{1}{\mathbb{P}[|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]}\right] - 1.$$

Lemma 5. *In Algorithm 1, we have*

$$\mathbb{E}\left[\sup_{\tau \geq \tau_q+1} \prod_{i \in Z} \frac{1}{\mathbb{P}[|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]}\right] - 1 \leq \begin{cases} \left(c\varepsilon^{-4}\right)^{|Z|} & \text{for every } q \geq 0 \\ e^{-\varepsilon^2 q/8}\left(c'\varepsilon^{-4}\right)^{|Z|} & \text{if } q > 8/\varepsilon^2, \end{cases}$$

where c and c' are two universal constants.

These lemmas allow us to get a constant regret under the event $\mathfrak{Z}_t \wedge \neg\mathfrak{C}_t$. Indeed, we have from Lemma 3 that

$$\begin{aligned}\sum_{t \in [T]} \mathbb{E}[\Delta_t \mathbb{I}\{\mathfrak{Z}_t \wedge \neg\mathfrak{C}_t\}] &\leq \Delta_{\max} \sum_{Z \subset A^*,\ Z \neq \emptyset} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{I}\{\mathfrak{S}_t(Z) \wedge \mathfrak{T}_t(Z)\}\right] \\ &= \Delta_{\max} \sum_{Z \subset A^*,\ Z \neq \emptyset} \sum_{q \geq 0} \mathbb{E}\left[\sum_{t=\tau_q+1}^{\tau_{q+1}} \mathbb{I}\{\mathfrak{S}_t(Z), \mathfrak{T}_t(Z)\}\right].\end{aligned}$$

Lemma 4 and 5 gives that the above is further upper bounded by

$$\Delta_{\max} \sum_{Z \subset A^*,\ Z \neq \emptyset} \left(\sum_{q=0}^{\lceil 8/\varepsilon^2 \rceil - 1} \left(c\varepsilon^{-4}\right)^{|Z|} + \sum_{q \geq \lceil 8/\varepsilon^2 \rceil} e^{-\varepsilon^2 q/8}\left(c'\varepsilon^{-4}\right)^{|Z|} \right)$$

which is bounded by

$$\Delta_{\max} \frac{C}{\varepsilon^2}\left(\frac{C'}{\varepsilon^4}\right)^{m^*},$$

where C and C' are two universal constants. This concludes the proof of the theorem.

Proof of Lemma 4. Since $\mathfrak{S}_t(Z), \mathfrak{T}_t(Z)$ are independent conditioned on the history $\mathcal{H}_t$, the LHS is

$$\mathbb{E}\left[\sum_{k \geq 1} (k-1)\mathbb{P}\left[\neg\mathfrak{T}_{t_{k,q}}(Z) \middle| \mathcal{H}_{t_{k,q}}\right] \prod_{j=1}^{k-1} \mathbb{P}\left[\mathfrak{T}_{t_{j,q}}(Z) \middle| \mathcal{H}_{t_{j,q}}\right]\right],$$

where $t_{k,q}$ is the round t where $\mathfrak{S}_t(Z)$ holds for the k-th time since the beginning of the round $\tau_q + 1$. Within the expectation, one can recognize the expectation of a time-varying geometric distribution, where the success probability of the k-th trial is $\mathbb{P}\left[\neg\mathfrak{T}_{t_{k,q}}(Z) \middle| \mathcal{H}_{t_{k,q}}\right]$. We can upper bound this inner expectation by the expectation of a geometric distribution whose success probability

$$\inf_{\tau \geq \tau_q+1} \mathbb{P}[\neg\mathfrak{T}_\tau(Z) | \mathcal{H}_\tau] = \inf_{\tau \geq \tau_q+1} \prod_{i \in Z} \mathbb{P}[|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]$$

is lower than all the success probabilities of the time-varying geometric distribution. This gives the result by monotonicity of the expectation, and rewriting the expectation of the geometric distribution. □

Proof of Lemma 5. For any arm $i \in [n]$, $k_i \in \mathbb{N}$, we define p_{i,k_i} as the probability of $\left|\widetilde{\theta}_{i,k_i} - \mu_i^*\right| \leq \varepsilon$, where $\widetilde{\theta}_{i,k_i}$ is a sample from the posterior of arm i when there are k_i observations of arm i (i.e., p_{i,k_i} is a random variable measurable with respect to those k_i independent draws of arm i). From Lemma 5,6 in Wang and Chen [2018], we know that

$$\mathbb{E}\left[\frac{1}{p_{i,k_i}}\right] \leq \begin{cases} 4/\varepsilon^2 & \text{for every } k_i \geq 0 \\ 1 + 6c'' \cdot e^{-\varepsilon^2 k_i/2}\varepsilon^{-2} + \frac{2}{e^{\varepsilon^2 k_i/8}-2} & \text{if } k_i > 8/\varepsilon^2, \end{cases}$$

for some universal constant c''. Since $\mathfrak{S}_t(Z) \wedge \neg\mathfrak{T}_t(Z)$ implies that $Z \subset A_t$, we know that for $\tau \geq \tau_q + 1$, $N_{i,\tau-1} \geq q$ for all $i \in Z$. Using the mutual independence of outcomes, and the fact that the distribution of $\theta_{i,\tau}$ depends only on the history of arm i, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\tau \geq \tau_q+1} \prod_{i \in Z} \frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]}\right] - 1 \\
&= \mathbb{E}\left[\sup_{\tau \geq \tau_q+1} \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]} - 1\right)\right] \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \mathbb{E}\left[\prod_{i \in Z'} \sup_{\tau \geq \tau_q+1} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \leq \varepsilon | \mathcal{H}_\tau]} - 1\right)\right] \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \mathbb{E}\left[\prod_{i \in Z'} \sum_{k_i \geq q} \left(\frac{1}{p_{i,k_i}} - 1\right)\right], \\
&= \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \mathbb{E}\left[\sum_{k_i \geq q} \left(\frac{1}{p_{i,k_i}} - 1\right)\right].
\end{aligned}$$

From this point, there are two cases: If $q > 8/\varepsilon^2$,

$$\begin{aligned}
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \sum_{k_i \geq q} \left(6c'' \cdot e^{-\varepsilon^2 k/2}\varepsilon^{-2} + 2e^{-\varepsilon^2 k/8}\left(1 - 2e^{-\varepsilon^2 k/8}\right)^{-1}\right) \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(6c'' \cdot e^{-\varepsilon^2 q/2}\varepsilon^{-2} \sum_{k \geq 0} e^{-\varepsilon^2 k/2} + 2e^{-\varepsilon^2 q/8}\left(1 - 2e^{-\varepsilon^2 q/8}\right)^{-1} \sum_{k \geq 0} e^{-\varepsilon^2 k/8}\right) \\
&= \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(6c'' \cdot e^{-\varepsilon^2 q/2}\varepsilon^{-2}\left(1 - e^{-\varepsilon^2/2}\right)^{-1} + 2e^{-\varepsilon^2 q/8}\left(1 - 2e^{-\varepsilon^2 q/8}\right)^{-1}\left(1 - e^{-\varepsilon^2/8}\right)^{-1}\right) \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(6c'' \cdot e^{-\varepsilon^2 q/2}\varepsilon^{-2} \cdot 2\varepsilon^{-2}\left(1 - e^{-1/2}\right)^{-1} + 2e^{-\varepsilon^2 q/8}\left(1 - 2e^{-1}\right)^{-1} \cdot 8\varepsilon^{-2}\left(1 - e^{-1/8}\right)^{-1}\right) \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} e^{-|Z'|\varepsilon^2 q/8}\left(12c'' \cdot e^{-3}\left(1 - e^{-1/2}\right)^{-1} \cdot \varepsilon^{-4} + 16\left(1 - 2e^{-1}\right)^{-1}\varepsilon^{-2}\left(1 - e^{-1/8}\right)^{-1}\right)^{|Z'|} \\
&\leq e^{-\varepsilon^2 q/8}\left(12c'' \cdot e^{-3}\left(1 - e^{-1/2}\right)^{-1}\varepsilon^{-4} + 16\left(1 - 2e^{-1}\right)^{-1}\varepsilon^{-2}\left(1 - e^{-1/8}\right)^{-1} + 1\right)^{|Z|} \\
&\leq e^{-\varepsilon^2 q/8}\left(c'\varepsilon^{-4}\right)^{|Z|},
\end{aligned}$$

and if $q \leq 8/\varepsilon^2$,

$$\begin{aligned}
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(\sum_{k=q}^{\lfloor 8/\varepsilon^2 \rfloor} (4/\varepsilon^2 - 1) + \sum_{k \geq \lfloor 8/\varepsilon^2 \rfloor + 1}^{\infty} \left(6c \cdot e^{-\varepsilon^2 k/2}\varepsilon^{-2} + 2e^{-\varepsilon^2 k/8}\left(1 - 2e^{-\varepsilon^2 k/8}\right)^{-1}\right)\right) \\
&\leq \sum_{Z' \subset Z,\ Z' \neq \emptyset} \prod_{i \in Z'} \left(36\varepsilon^{-4} + 12c \cdot e^{-4}\left(1 - e^{-1/2}\right)^{-1}\varepsilon^{-4} + 16e^{-1}\left(1 - 2e^{-1}\right)^{-1}\varepsilon^{-2}\left(1 - e^{-1/8}\right)^{-1}\right)
\end{aligned}$$

$\leq \left(c\varepsilon^{-4}\right)^{|Z|}$,

where c, c' are two universal constant. □

A.3 Discussion on the new exponential constant term (step 4 in the above proof)

We give here an explanation concerning the modification of Lemma 7 from Wang and Chen [2018]. First, we respectfully disagree with the end of their proof, where the expected number of time slots for $\mathfrak{S}_t(Z) \wedge \neg\mathfrak{T}_t(Z)$ to occur is a weighted mean of expectations where the counters are fixed and non-random. To obtain such a weighted mean, they have conditioned on the value of the counters. However, counters depend on the chosen action, and thus on the outcomes previously obtained, so conditioning on it would modify the expectation, since the term inside the expectation not only depends on counters, but also on outcomes obtained so far. To illustrate more clearly this point, let us focus on one arm i, and consider the extreme case where we get a new sample (i.e. the counter is incremented) only if samples $Y_{i,t}$ previously obtained from i were all 0, say. Then conditioning on the fact that the counter is incremented would remove all the randomness of samples $Y_{i,t}$, and we thus can't consider an expectation on those samples as if their randomness was not impacted.

We now expose our approach to overcome this issue. We first rewrite the above mentioned expected number of time slots as the expectation (over the history) of the expectation of a time-varying geometric distribution, where the time-varying success probability depends on the history. The inner expectation can be bounded by the expectation of a geometric distribution whose success probability is the infimum over all the success probabilities of the time-varying geometric distribution. Let's note that this gives us the inverse success probability minus one, as in Wang and Chen [2018], but that counters are still random. We use that this inverse probability can be factorized: from the relation $\prod_{i\in A} a_i - 1 = \sum_{A'\subset A,\ A'\neq\emptyset} \prod_{i\in A'}(a_i - 1)$, valid for any vector $\mathbf{a} = (a_i)$ on a set A, and from the mutual independence of outcomes, we're reduced to bounding the expectation in the one-dimensional case. To overcome the randomness of the counters, we use an union bound. It is this union bound that brings a larger dependence on the constant term, because it forces us to look at a sum of the form $\sum_q \sum_{k\geq q} x_k$, instead of a simply $\sum_q x_q$. Let's remark that Wang and Chen [2018] use the eventual exponential decreasing of the sequence (x_q) in order to get their final bound. We manage to deal with the sequence $\left(\sum_{k\geq q} x_k\right)$ instead, by noticing that the eventual exponential decreasing of the sequence (x_q) implies the eventual exponential decreasing of the sequence $\left(\sum_{k\geq q} x_k\right)$.

B Proof of Proposition 1

Assumption 4 encompasses κ_i^2-sub Gaussian outcomes with $D_i = \kappa_i^2 m$ for all $i \in [n]$. Indeed, let $\boldsymbol{\lambda} = \boldsymbol{\lambda} \odot \mathbf{e}_A$ for some action A and observe that

$$\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq \mathbb{E}\left[\sum_i \frac{|\kappa_i\lambda_i|}{\|\boldsymbol{\kappa}\odot\boldsymbol{\lambda}\|_1} e^{\|\boldsymbol{\kappa}\odot\boldsymbol{\lambda}\|_1 \operatorname{sign}(\lambda_i)\frac{X_i-\mu_i^*}{\kappa_i}}\right] \leq e^{\|\boldsymbol{\kappa}\odot\boldsymbol{\lambda}\|_1^2/2} \leq e^{\|\boldsymbol{\kappa}\odot\boldsymbol{\lambda}\|_2^2|A|/2} \leq e^{\|\boldsymbol{\kappa}\odot\boldsymbol{\lambda}\|_2^2 m/2}.$$

The case of $\mathbf{C}$-sub-Gaussian outcomes with a known sub-Gaussian matrix $\mathbf{C}$ (i.e., $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal\mathbf{C}\boldsymbol{\lambda}/2}$ for all $\boldsymbol{\lambda}\in\mathbb{R}^n$) is also captured, taking[5] $D_i = \max_{A\in\mathcal{A},\ i\in A}\sum_{j\in A}|C_{ij}|$. Indeed, for an action A,

$$\sum_{i,j\in A}\lambda_i\lambda_j C_{ij} \leq \sum_{i,j\in A}\frac{\lambda_i^2+\lambda_j^2}{2}|C_{ij}| = \sum_{i\in A}\lambda_i^2\sum_{j\in A}|C_{ij}| \leq \sum_{i\in n}\lambda_i^2 \max_{A\in\mathcal{A},\ i\in A}\sum_{j\in A}|C_{ij}|.$$

C Proof of Theorem 2

We beginning by stating the complete version of Theorem 2.

[5] This D_i can be computed whenever linear maximization on $\mathcal{A}$ is efficient: for x high enough, we have $\max_{A\in\mathcal{A},\ i\in A}\sum_{j\in A}|C_{ij}| = C_{ii} - x + \max_{A\in\mathcal{A}}\sum_{j\in A}(|C_{ij}|\mathbb{I}\{j\neq i\} + x\mathbb{I}\{j=i\})$.

Theorem. *The policy π described in Algorithm 2 has regret $R_T(\pi)$ bounded by*

$$256\log_2^2(4\sqrt{m})\sum_{i\in[n]}\frac{B^2\beta D_i\log(2^m|\mathcal{A}|T)}{\Delta_{i,\min}}+\Delta_{\max}(1+2n)$$
$$+\frac{nm^2\Delta_{\max}}{\left(\frac{\Delta_{\min}}{2B}-(m^{*2}+1)\varepsilon\right)^2}+\Delta_{\max}\left(C\varepsilon^{-2}\beta\max_i D_i\right)\left(\frac{C'}{\sqrt{\beta-1}}\varepsilon^{-4}\beta^3\max_i D_i^2\right)^{m^*},$$

where C, C' are two universal constants, and $\varepsilon\in(0,1)$ is such that $\Delta_{\min}/(2B)-(m^{*2}+1)\varepsilon>0$.

For the proof of Theorem 2, we consider the same events as in the proof of Theorem 1, except for the event $\mathfrak{D}_t$, that becomes

$$\mathfrak{D}_t\triangleq\left\{\left\|\mathbf{e}_{A_t}\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\sqrt{2\log(|\mathcal{A}|2^mT)\sum_{i\in A_t}\beta D_i/N_{i,t-1}}\right\}.$$

Step 1 is unchanged. Step 2 and Step 3 are modified only through the event $\mathfrak{D}_t$, using the following modification of Lemma 2.

Lemma 6. *In Algorithm 2, for all round t, we have that $\mathbb{P}[\mathfrak{D}_t|\mathcal{H}_t]\leq 1/T$.*

Proof. We rely on the fact that conditionally on the history, the sample $\boldsymbol{\theta}_t$ is Gaussian of mean $\overline{\boldsymbol{\mu}}_{t-1}$ and of diagonal covariance given by $\beta D_i N_{i,t-1}^{-1}$. We thus define the functions

$$\alpha_t(A)\triangleq\sqrt{2\log(|\mathcal{A}|2^mT)\sum_{i\in A}\frac{\beta D_i}{N_{i,t-1}}},\quad\text{and}\quad\lambda_t(A)\triangleq\frac{\alpha_t(A)}{\sum_{i\in A}\beta D_i/N_{i,t-1}},$$

we have

$$\begin{aligned}
\mathbb{P}\left[\left\|\mathbf{e}_{A_t}\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\alpha_t(A_t)\middle|\mathcal{H}_t\right]&\leq\sum_{A\in\mathcal{A}}\mathbb{P}\left[\left\|\mathbf{e}_A\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1\geq\alpha_t(A)\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\mathbb{E}\left[e^{\lambda_t(A)\left\|\mathbf{e}_A\odot\left(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1}\right)\right\|_1}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\prod_{i\in A}\mathbb{E}\left[e^{\lambda_t(A)\left|\theta_{i,t}-\overline{\mu}_{i,t-1}\right|}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}e^{-\lambda_t(A)\alpha_t(A)}\prod_{i\in A}\mathbb{E}\left[e^{\lambda_t(A)\left(\theta_{i,t}-\overline{\mu}_{i,t-1}\right)}+e^{\lambda_t(A)\left(\overline{\mu}_{i,t-1}-\theta_{i,t}\right)}\middle|\mathcal{H}_t\right]\\
&\leq\sum_{A\in\mathcal{A}}2^{|A|}e^{-\lambda_t(A)\alpha_t(A)}e^{\lambda_t(A)^2\sum_{i\in A}\beta D_i/(2N_{i,t-1})}\leq 1/T.
\end{aligned}$$

□

The final bound on the regret in Step 3 is obtained using the same derivation as in Theorem 1, which gives the following leading term:

$$256\log_2^2(4\sqrt{m})\sum_{i\in[n]}\frac{B^2\beta D_i\log(2^m|\mathcal{A}|T)}{\Delta_{i,\min}}.$$

In the following, we consider the last step, consisting in bounding the regret under the event $\mathfrak{Z}_t$ and $\neg\mathfrak{C}_t$. From the initialization phase, we also assume that the event

$$\mathfrak{M}_t\triangleq\{\forall i\in[n],\ N_{i,t-1}\geq 1\}$$

holds (the regret under the complementary event is clearly bounded by $n\Delta_{\max}$). If there is no initialization, we can have $q=0$ in the following, noticing that when $\theta_{i,t}$ is uniform on $[a,b]$, then the probability $\mathbb{P}[|\theta_{i,t}-\mu_i^*|\leq\varepsilon|\mathcal{H}_t]$ is equal to $2\varepsilon/(b-a)$.

Step 4: bound under $\mathfrak{M}_t \wedge \mathfrak{Z}_t \wedge \neg\mathfrak{C}_t$ We use the independence of the prior, as for Theorem 1, to obtain the following upper bound, using $\mathfrak{M}_t$ to be able to start from $q = 1$.

$$\sum_{t\in[T]} \mathbb{E}[\Delta(A_t)\mathbb{I}\{\mathfrak{M}_t \wedge \mathfrak{Z}_t \wedge \neg\mathfrak{C}_t\}] \leq \sum_{Z\subset A^*,\, Z\neq\emptyset} \sum_{q\geq 1} \mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \sum_{Z'\subset Z,\, Z'\neq\emptyset} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau}-\mu_i^*|\leq\varepsilon|\mathcal{H}_\tau]} - 1\right)\right]$$
$$\leq \sum_{Z\subset A^*,\, Z\neq\emptyset} \sum_{q\geq 1} \underbrace{\sum_{Z'\subset Z,\, Z'\neq\emptyset} \mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau}-\mu_i^*|\leq\varepsilon|\mathcal{H}_\tau]} - 1\right)\right]}_{(4)}.$$

However, the expectation can't be put inside the product since outcomes are not mutually independent. We can still take a union bound on counters:

$$(4) \leq \sum_{Z'\subset Z,\, Z'\neq\emptyset} \sum_{\mathbf{k}\in[q..\infty)^{Z'}} \mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \mathbb{I}\{\forall i\in Z',\ N_{i,\tau-1}=k_i\} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau}-\mu_i^*|\leq\varepsilon|\mathcal{H}_\tau]} - 1\right)\right].$$

One can notice that for all $i \in Z'$, all $k_i \geq q$, $\mathbb{I}\{N_{i,\tau-1}=k_i\}\left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau}-\mu_i^*|\leq\varepsilon|\mathcal{H}_\tau]} - 1\right)$ is of the form $\mathbb{I}\{N_{i,\tau-1}=k_i\}g_i\big(|\overline{\mu}_{i,\tau-1}-\mu_i^*|\big)$, with g_i being an increasing function on $\mathbb{R}_+$. Indeed, we see that the conditional distribution of $\theta_{i,\tau}-\overline{\mu}_{i,\tau-1}$ is $\mathcal{N}\big(0,\beta D_i N_{i,\tau-1}^{-1}\big)$, which is symmetric, so we have

$$\mathbb{P}[\,|\theta_{i,\tau}-\mu_i^*|\leq\varepsilon|\mathcal{H}_\tau] = \mathbb{P}\big[\,\big|\theta_{i,\tau}-\overline{\mu}_{i,\tau-1}+\big|\overline{\mu}_{i,\tau-1}-\mu_i^*\big|\big|\leq\varepsilon\big|\mathcal{H}_\tau\big].$$

In addition, under $\mathbb{I}\{N_{i,\tau-1}=k_i\}$, the conditional distribution of $\theta_{i,\tau}-\overline{\mu}_{i,\tau-1}$ does not depend on the history, but only on k_i. Therefore, the above probability is a function of $\big|\overline{\mu}_{i,\tau-1}-\mu_i^*\big|$ and so the function g_i exists. It is increasing on $\mathbb{R}_+$ because for any fixed $\sigma > 0$,

$$\frac{\partial}{\partial x}\int_{x-\varepsilon}^{x+\varepsilon} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2}{2\sigma^2}}\mathrm{d}u = \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{(x+\varepsilon)^2}{2\sigma^2}} - e^{-\frac{(x-\varepsilon)^2}{2\sigma^2}}\right) < 0 \text{ for } x > 0.$$

In particular, we can consider the inverse function g_i^{-1}. We now want to use a stochastic dominance argument in order to treat the outcomes as if they were Gaussian: we have for any $\mathbf{k}\in[q..\infty)^{Z'}$,

$$\mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \prod_{i\in Z'} \big(\mathbb{I}\{N_{i,\tau-1}=k_i\}g_i\big(|\overline{\mu}_{i,\tau-1}-\mu_i^*|\big)\big)\right]$$
$$= \mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \prod_{i\in Z'} \left(\mathbb{I}\{N_{i,\tau-1}=k_i\}\int_0^\infty \mathbb{I}\big\{g_i\big(|\overline{\mu}_{i,\tau-1}-\mu_i^*|\big)\geq u_i\big\}\mathrm{d}u_i\right)\right]$$
$$\leq \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \mathbb{E}\left[\sup_{\tau\geq\tau_q+1} \prod_{i\in Z'} \mathbb{I}\{N_{i,\tau-1}=k_i\}\mathbb{I}\big\{g_i\big(|\overline{\mu}_{i,\tau-1}-\mu_i^*|\big)\geq u_i\big\}\right]\mathrm{d}\mathbf{u}$$
$$= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \mathbb{E}\left[\prod_{i\in Z'} \mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\big\{g_i\big(|\overline{\mu}_{i,\tau^*-1}-\mu_i^*|\big)\geq u_i\big\}\right]\mathrm{d}\mathbf{u}, \tag{5}$$

where τ^* is the first time τ such that the event $\mathbb{I}\big\{\forall i\in Z',\ N_{i,\tau-1}=k_i$ and $g_i\big(|\overline{\mu}_{i,\tau-1}-\mu_i^*|\big)\geq u_i\big\}$ holds, and is ∞ if it never holds.

$$(5) = \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \mathbb{E}\left[\prod_{i\in Z'} \mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\big\{g_i\big(|\overline{\mu}_{i,\tau^*-1}-\mu_i^*|\big)\geq u_i\vee g_i(0)\big\}\right]\mathrm{d}\mathbf{u}$$
$$= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \mathbb{E}\left[\prod_{i\in Z'} \mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\big\{|\overline{\mu}_{i,\tau^*-1}-\mu_i^*|\geq g_i^{-1}(u_i\vee g_i(0))\big\}\right]\mathrm{d}\mathbf{u}$$
$$= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \sum_{\mathbf{s}\in\{-1,1\}^{Z'}} \underbrace{\mathbb{E}\left[\prod_{i\in Z'} \mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\big\{s_i(\overline{\mu}_{i,\tau^*-1}-\mu_i^*)\geq g_i^{-1}(u_i\vee g_i(0))\big\}\right]}_{(6)}\mathrm{d}\mathbf{u}$$

$$
\begin{aligned}
(6) &\leq \mathbb{P}\left[\frac{\exp\left(\sum_{i\in Z'} N_{i,\tau^*-1}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}\left(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\right)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)}{\exp\left(\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right)}\geq 1, (N_{i,\tau^*-1})_{i\in Z'}=\mathbf{k}\right]\\
&\leq \mathbb{P}\left[\frac{\exp\left(\sum_{i\in Z'} N_{i,\tau^*-1}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}\left(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\right)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)}{\exp\left(\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right)}\geq 1\right]\\
&\leq \frac{\mathbb{E}\left[\exp\left(\sum_{i\in Z'} N_{i,\tau^*-1}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}\left(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\right)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)\right]}{\exp\left(\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right)}\\
&= \frac{\mathbb{E}\left[\exp\left(\sum_{t=1}^{\tau^*-1}\sum_{i\in Z'\cap A_t}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}(X_{i,t}-\mu_i^*)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)\right]}{\exp\left(\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right)}.
\end{aligned}
$$

From Assumption 4, we have that

$$
M_\tau = \exp\left(\sum_{t=1}^{\tau-1}\sum_{i\in Z'\cap A_t}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}(X_{i,t}-\mu_i^*)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)
$$

is a supermartingale:

$$
\begin{aligned}
\mathbb{E}[M_\tau|\mathcal{F}_{\tau-1}] &= M_{\tau-1}\mathbb{E}\left[\exp\left(\sum_{i\in Z'\cap A_{\tau-1}}\left(\frac{s_i g_i^{-1}(u_i\vee g_i(0))}{D_i}(X_{i,\tau-1}-\mu_i^*)-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2}{2D_i}\right)\right)\middle|\mathcal{F}_{\tau-1}\right]\\
&\leq M_{\tau-1}.
\end{aligned}
$$

Since τ^* is a stopping time with respect to $\mathcal{F}_\tau$, we have from Doob's optional sampling theorem for non-negative supermartingales[6] that $\mathbb{E}[M_{\tau^*}]\leq 1$. Therefore,

$$
(6)\leq \exp\left(-\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right).
$$

Now, we want to use the following fact (see Chang et al. [2011]): if $\eta\sim\mathcal{N}(0,1)$, then with $\beta>1$,

$$
\sqrt{\frac{2e}{\pi}}\frac{\sqrt{\beta-1}}{\beta}e^{-\beta x^2/2}\leq \mathbb{P}[|\eta|\geq x].
$$

Indeed, this gives

$$
\sqrt{\frac{2e}{\pi}}\frac{\sqrt{\beta-1}}{\beta}\exp\left(-\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right)\leq \mathbb{P}\left[|\eta_i|\geq g_i^{-1}(u_i\vee g_i(0))\sqrt{\frac{k_i}{\beta D_i}}\right],
$$

where $\boldsymbol{\eta}\sim\mathcal{N}(0,1)^{\otimes Z'}$. Thus,

$$
\begin{aligned}
(5) &\leq \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\prod_{i\in Z'}\mathbb{P}\left[\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\geq g_i^{-1}(u_i\vee g_i(0))\right]\mathrm{d}\mathbf{u}\\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\prod_{i\in Z'}\mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\geq u_i\vee g_i(0)\right]\mathrm{d}\mathbf{u}
\end{aligned}
$$

[6] We use the version that relies on Fatou's lemma (Durrett [2019], Theorem 5.7.6), so that it is not needed to have any additional condition on the stopping time τ^*.

$$
\begin{aligned}
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|} \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}} \prod_{i\in Z'} \mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right) \geq u_i\right] \mathrm{d}\mathbf{u} \\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|} \prod_{i\in Z'} \int_0^{\infty} \mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right) \geq u_i\right] \mathrm{d}u_i \\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|} \prod_{i\in Z'} \mathbb{E}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\right].
\end{aligned}
$$

We now want to bound $\mathbb{E}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\right]$. We define $\alpha = 2-\sqrt{2}$, the unique solution in $(1/2,1)$ of $\alpha - 1/2 = (\alpha-1)^2/2$. Notice that $\alpha - 1/2 \geq 1/12$. Define $\varepsilon_i \triangleq \varepsilon\sqrt{\frac{k_i}{\beta D_i}}$. By definition, we have

$$
\begin{aligned}
\mathbb{E}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\right] &= \int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{\int_{x-\varepsilon_i}^{x+\varepsilon_i} e^{-y^2/2}\mathrm{d}y}\mathrm{d}x - 1 \\
&= \underbrace{2\int_{\alpha\varepsilon_i}^{+\infty} \frac{1}{\int_{x-\varepsilon_i}^{x+\varepsilon_i} e^{-\frac{y^2-x^2}{2}}\mathrm{d}y}\mathrm{d}x}_{A_1} + \underbrace{\int_{-\alpha\varepsilon_i}^{\alpha\varepsilon_i} \frac{e^{-x^2/2}}{\int_{x-\varepsilon_i}^{x+\varepsilon_i} e^{-y^2/2}\mathrm{d}y}\mathrm{d}x}_{A_2} - 1\,.
\end{aligned}
$$

We first bound A_1. With the change of variable $u = y - x$, we get:

$$
\begin{aligned}
A_1 &= 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{1}{\int_{-\varepsilon_i}^{\varepsilon_i} e^{-u^2/2-ux}\mathrm{d}u}\mathrm{d}x \\
&\leq 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{1}{\int_{-\varepsilon_i}^{0} e^{-u^2/2-ux}\mathrm{d}u}\mathrm{d}x
\end{aligned}
$$

Note that for $x \geq \alpha\varepsilon_i$ and $u \in [-\varepsilon_i, 0]$, $-u^2/2 - ux \geq -(1-\frac{1}{2\alpha})ux$ and thus:

$$
\begin{aligned}
A_1 &\leq 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{1}{\int_{-\varepsilon_i}^{0} e^{-(1-\frac{1}{2\alpha})ux}\mathrm{d}u}\mathrm{d}x \\
&= 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{(1-\frac{1}{2\alpha})x}{e^{(1-\frac{1}{2\alpha})\varepsilon_i x} - 1}\mathrm{d}x. \qquad (7)
\end{aligned}
$$

We distinguish two regimes. First, if $\varepsilon_i^2 \geq 12$, then

$$
\begin{aligned}
(7) &\leq \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1} \int_{\alpha\varepsilon_i}^{+\infty} \left(1-\frac{1}{2\alpha}\right) x e^{-(1-\frac{1}{2\alpha})\varepsilon_i x}\mathrm{d}x \\
&= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1} \frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2} \int_{(\alpha-\frac{1}{2})\varepsilon_i^2}^{+\infty} x e^{-x}\mathrm{d}x \\
&= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1} \frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2} \left[-(x+1)e^{-x}\right]_{(\alpha-\frac{1}{2})\varepsilon_i^2}^{\infty} \\
&= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1} \frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2} \left(\left(\alpha-\frac{1}{2}\right)\varepsilon_i^2 + 1\right) e^{-(\alpha-\frac{1}{2})\varepsilon_i^2} \\
&= \frac{2}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1} \left(\alpha + \frac{\alpha}{(\alpha-\frac{1}{2})\varepsilon_i^2}\right) \\
&\leq 4e^{-\varepsilon_i^2/12}.
\end{aligned}
$$

Otherwise, we have

$$
(7) = \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2} \int_{\alpha\varepsilon_i^2}^{\infty} \frac{u}{e^{(1-\frac{1}{2\alpha})u} - 1}\mathrm{d}u
$$

$$
\begin{aligned}
&\leq \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2}\int_0^\infty \frac{u}{e^{(1-\frac{1}{2\alpha})u}-1}\mathrm{d}u \\
&= \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2}\frac{\pi^2}{6(1-\frac{1}{2\alpha})^2} \\
&\leq \frac{24\beta D_i}{\varepsilon^2}.
\end{aligned}
$$

We now bound A_2. As $x \in [-\alpha\varepsilon_i, \alpha\varepsilon_i]$, it comes that $[-(1-\alpha)\varepsilon_i, (1-\alpha)\varepsilon_i] \subset [x-\varepsilon_i, x+\varepsilon_i]$. This implies that

$$
\begin{aligned}
A_2 &\leq \frac{\int_{-\alpha\varepsilon_i}^{\alpha\varepsilon_i} e^{-x^2/2}\mathrm{d}x}{\int_{-(1-\alpha)\varepsilon_i}^{(1-\alpha)\varepsilon_i} e^{-x^2/2}\mathrm{d}x} - 1 \\
&= \frac{2\int_{(1-\alpha)\varepsilon_i}^{\alpha\varepsilon_i} e^{-x^2/2}\mathrm{d}x}{\int_{-(1-\alpha)\varepsilon_i}^{(1-\alpha)\varepsilon_i} e^{-x^2/2}\mathrm{d}x} \\
&\leq \frac{2\int_{(1-\alpha)\varepsilon_i}^{\infty} e^{-x^2/2}\mathrm{d}x}{\int_{-(1-\alpha)\varepsilon_i}^{(1-\alpha)\varepsilon_i} e^{-x^2/2}\mathrm{d}x} \\
&\leq \frac{e^{-(1-\alpha)^2\varepsilon_i^2/2}}{1-e^{-(1-\alpha)^2\varepsilon_i^2/2}} \leq \left(1+\frac{12}{\varepsilon_i^2}\right)e^{-\varepsilon_i^2/12}.
\end{aligned}
$$

The penultimate inequality relies on $\int_x^\infty e^{-u^2/2}\mathrm{d}u \leq \sqrt{\frac{\pi}{2}}e^{-x^2/2}$ (see Jacobs and Wozencraft [1965], eq. (2.122)). We obtain again two regimes: $2e^{-\varepsilon_i^2/12}$ if $\varepsilon_i^2 \geq 12$, and $1+\frac{12\beta D_i}{\varepsilon^2}$ otherwise. To summarize, we proved that

$$
(5) \leq \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\prod_{i\in Z'}\left(\mathbb{I}\left\{\varepsilon^2\frac{k_i}{\beta D_i}<12\right\}\left(1+36\frac{\beta D_i}{\varepsilon^2}\right)+\mathbb{I}\left\{\varepsilon^2\frac{k_i}{\beta D_i}\geq 12\right\}6e^{-\varepsilon^2\frac{k_i}{12\beta D_i}}\right).
$$

After the summation on $\mathbf{k}$, on Z', on q, and on Z, we obtain that there exists two constants C, C' such that

$$
\sum_{Z\subset A^*,\ Z\neq\emptyset}\ \sum_{q\geq 1}\ \sum_{Z'\subset Z,\ Z'\neq\emptyset}\ \sum_{\mathbf{k}\in[q..\infty)^{Z'}} (5) \leq \left(C\varepsilon^{-2}\beta\max_i D_i\right)\left(\frac{C'\beta}{\sqrt{\beta-1}}\varepsilon^{-4}\beta^2\max_i D_i^2\right)^{m^*}.
$$

Thus,

$$
\sum_{t\in[T]}\mathbb{E}[\Delta(A_t)\mathbb{I}\{\mathfrak{M}_t\wedge\mathfrak{Z}_t\wedge\neg\mathfrak{C}_t\}] \leq \Delta_{\max}\left(C\varepsilon^{-2}\beta\max_i D_i\right)\left(\frac{C'\beta}{\sqrt{\beta-1}}\varepsilon^{-4}\beta^2\max_i D_i^2\right)^{m^*}.
$$

D Proof of Theorem 3 (CLIP CTS-GAUSSIAN for linear rewards)

In this section, we provide an analysis for the regret bound of CLIP CTS-GAUSSIAN, which is stated completely as follows.

Theorem. *The policy* CLIP CTS-GAUSSIAN *has regret bounded by*

$$
\sum_{i\in[n]}\frac{128\left(4\log_2^2(4\sqrt{m})\beta D_i\log(2^m|\mathcal{A}|T)\wedge m\Gamma_{ii}(\log(T)+4\log\log(T))\right)}{\Delta_{i,\min}}+\Delta_{\max}(1+5.2n)
$$

$$
+\frac{nm^2\Delta_{\max}}{\left(\frac{\Delta_{\min}}{2B}-(m^*(m^*+1)/2+1)\varepsilon\right)^2}+\Delta_{\max}\left(C\varepsilon^{-2}\beta\max_i D_i\right)\left(\frac{C'}{\sqrt{\beta-1}}\varepsilon^{-4}\beta^3\max_i D_i^2\right)^{m^*},
$$

where C, C' are two universal constants, and $\varepsilon \in (0,1)$ is such that $\Delta_{\min}/(2B) - ({m^*}^2+1)\varepsilon > 0$.

More precisely, notice that the modification on the sample $\boldsymbol{\theta}_t$ has an impact only in two places in the analysis: in the concentration bound and in the event controlling optimism. We detail these two points in the following.

D.1 Concentration bound

In this subsection, we provide the concentration bound of CLIP CTS-GAUSSIAN. Our strategy here is to either use the concentration from $\boldsymbol{\mu}_t$ or from $\boldsymbol{\theta}_t$, depending on which regime is the best for each arm. Thus, we define $S \triangleq \left\{i \in [n],\ \Gamma_{ii} m(\log(T) + 4\log\log(T)) \geq 4\log_2^2(4\sqrt{m})\beta D_i \log(|A|2^m T)\right\}$. We have the following lemma.

Lemma 7.

$$\mathbb{P}\left[\mathbf{e}_{A_t\cap S}^\intercal\left(\overline{\boldsymbol{\mu}}_{t-1} \vee \boldsymbol{\theta}_t \wedge \boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1}\right) \geq \sqrt{2\log(|\mathcal{A}|2^m T)\sum_{i\in A_t\cap S} \beta D_i / N_{i,t-1}}\middle| \mathcal{H}_t\right] \leq 1/T.$$

Proof. We define the functions

$$\alpha_t(A) \triangleq \sqrt{2\log(|\mathcal{A}|2^m T)\sum_{i\in A}\frac{\beta D_i}{N_{i,t-1}}}, \quad \text{and} \quad \lambda_t(A) \triangleq \frac{\alpha_t(A)}{\sum_{i\in A}\beta D_i / N_{i,t-1}},$$

we have

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{e}_{A_t\cap S}^\intercal\left(\overline{\boldsymbol{\mu}}_{t-1} \vee \boldsymbol{\theta}_t \wedge \boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1}\right) \geq \alpha_t(A_t\cap S)\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} \mathbb{P}\left[\mathbf{e}_{A\cap S}^\intercal\left(\overline{\boldsymbol{\mu}}_{t-1} \vee \boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right) \geq \alpha_t(A\cap S)\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} e^{-\lambda_t(A\cap S)\alpha_t(A\cap S)}\mathbb{E}\left[e^{\lambda_t(A\cap S)\left\|\mathbf{e}_{A\cap S}\odot\left(0\vee\left(\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1}\right)\right)\right\|_1}\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} e^{-\lambda_t(A\cap S)\alpha_t(A\cap S)}\prod_{i\in A\cap S}\mathbb{E}\left[e^{\lambda_t(A\cap S)\left(0\vee\left(\theta_{i,t} - \overline{\mu}_{i,t-1}\right)\right)}\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} e^{-\lambda_t(A\cap S)\alpha_t(A\cap S)}\prod_{i\in A\cap S}\mathbb{E}\left[1 + e^{\lambda_t(A\cap S)\left(\theta_{i,t} - \overline{\mu}_{i,t-1}\right)}\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} e^{-\lambda_t(A\cap S)\alpha_t(A\cap S)}\prod_{i\in A\cap S}\mathbb{E}\left[2e^{\lambda_t(A\cap S)\left(\theta_{i,t} - \overline{\mu}_{i,t-1}\right)}\middle|\mathcal{H}_t\right] \\
&\leq \sum_{A\in\mathcal{A}} 2^{|A\cap S|} e^{-\lambda_t(A\cap S)\alpha_t(A\cap S)} e^{\lambda_t(A\cap S)^2\sum_{i\in A\cap S}\beta D_i/(2N_{i,t-1})} \\
&\leq 1/T.
\end{aligned}$$

□

We now use the definition of $\boldsymbol{\mu}_t$ to have

$$\mathbf{e}_{A_t\cap S^c}^\intercal\left(\overline{\boldsymbol{\mu}}_{t-1} \vee \boldsymbol{\theta}_t \wedge \boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1}\right) \leq \mathbf{e}_{A_t\cap S^c}^\intercal\left(\boldsymbol{\mu}_t - \overline{\boldsymbol{\mu}}_{t-1}\right) = \sum_{i\in A_t\cap S^c}\sqrt{\Gamma_{ii}\frac{2(\log(t) + 4\log\log(t))}{N_{i,t-1}}}.$$

To conclude, we have the following event

$$\mathfrak{A}_t \triangleq \left\{\Delta_t \leq \sqrt{8\log(|\mathcal{A}|2^m T)\sum_{i\in A_t\cap S}\beta D_i/N_{i,t-1}} + \sum_{i\in A_t\cap S^c}\sqrt{\Gamma_{ii}\frac{8(\log(t) + 4\log\log(t))}{N_{i,t-1}}}\right\}.$$

Using Proposition 4, we have

$$\begin{aligned}
\sum_{t\in[T]}\mathbb{E}[\Delta_t\mathbb{I}\{\mathfrak{A}_t\}] \leq \sum_{t\in[T]}\mathbb{E}&\left[\Delta_t\mathbb{I}\left\{\Delta_t \leq 2\sqrt{8\log(|\mathcal{A}|2^m T)\sum_{i\in A_t\cap S}\beta D_i/N_{i,t-1}}\right\}\right] \\
&+ \sum_{t\in[T]}\mathbb{E}\left[\Delta_t\mathbb{I}\left\{\Delta_t \leq 2\sum_{i\in A_t\cap S^c}\sqrt{\Gamma_{ii}\frac{8(\log(t) + 4\log\log(t))}{N_{i,t-1}}}\right\}\right].
\end{aligned}$$

We can thus apply Theorem 5 and Theorem 4 (see Appendix E) to get the bound

$$512\log_2^2(4\sqrt{m})\sum_{i\in S}\Delta_{i,\min}^{-1}\beta D_i\log(|\mathcal{A}|2^m T) + 128m\sum_{i\in S^c}\Delta_{i,\min}^{-1}\Gamma_{ii}(\log(T) + 4\log\log(T))$$

D.2 Optimism

In this subsection, we examine the theoretical impact of considering CLIP CTS-GAUSSIAN on the optimism-controlling event (event $\neg\mathfrak{C}_t$), in the case of linear rewards. For this purpose, we modify the beginning of Step 4 in the analysis by considering the following events.

- $\mathfrak{Z}_t \triangleq \{\Delta_t > 0\}$
- $\mathfrak{C}_t \triangleq \left\{\mathbf{e}_{A_t}^\intercal \widetilde{\boldsymbol{\theta}}_t > \mathbf{e}_{A^*}^\intercal \boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon\right\}$
- $\mathfrak{R}(\boldsymbol{\theta}', Z) \triangleq \left\{\forall A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal(\boldsymbol{\theta}') \text{ we have } Z \subset A,\ \mathbf{e}_{\mathrm{Oracle}(\boldsymbol{\theta}')}^\intercal \boldsymbol{\theta}' > \mathbf{e}_{A^*}^\intercal \boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon\right\}$
- $\mathfrak{S}_t(Z) \triangleq \left\{\forall \boldsymbol{\theta}' \text{ s.t. } 0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_Z \leq \varepsilon \mathbf{e}_Z,\ \mathfrak{R}(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}, Z) \text{ holds}\right\}$
- $\mathfrak{T}_t(Z) \triangleq \left\{\exists i \in Z,\ \mu_i^* - \mu_i^* \wedge \widetilde{\theta}_{i,t} > \varepsilon\right\}$.
- $\mathfrak{J}_t \triangleq \{\forall i \in [n], \mu_i^* \leq \mu_{i,t}\}$

In the above events, $\widetilde{\boldsymbol{\theta}}_t$ is $\boldsymbol{\mu}_t \wedge \boldsymbol{\theta}_t \vee \overline{\boldsymbol{\mu}}_t$. The last event $\mathfrak{J}_t$ holds with probability at least $1 - n/(t\log^2(t))$ from Hoeffding's inequality [Hoeffding, 1963]. We thus assume that this event hods in the following, since the regret under the complementary event is bounded by $3.2n\Delta_{\max}$. We first state the following lemma.

Lemma 8.

$$\mathfrak{Z}_t, \neg\mathfrak{C}_t \Rightarrow \exists Z \subset A^*,\ Z \neq \emptyset \textit{ s.t. the event } \mathfrak{S}_t(Z) \wedge \mathfrak{T}_t(Z) \textit{ holds.}$$

This allows us to consider the success probability $\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]$ in the analysis. Notice however that $Z \subset \mathrm{Oracle}\left(\left(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\right) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)$, that is guaranteed when $\mathfrak{S}_t(Z) \wedge \neg\mathfrak{T}_t(Z)$ holds, does not necessarily implies that $Z \subset \mathrm{Oracle}\left(\widetilde{\boldsymbol{\theta}}_t\right)$. However, it turns out that we have $Z \subset A$ for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal \left(\left(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\right) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)$ implies that $Z \subset A$ for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal \left(\widetilde{\boldsymbol{\theta}}_t\right)$. This last fact is from Lemma 9, with $\boldsymbol{\eta} = \left(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\right) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}$ and $\boldsymbol{\delta} = \left(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\right) \odot \mathbf{e}_Z$.

Lemma 9. *Let $\boldsymbol{\eta} \in \mathbb{R}^n$, $\boldsymbol{\delta} \in \mathbb{R}_+^n$ such that for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal \boldsymbol{\eta}$, we have $Z \subset A$. Then, for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z)$, we have $Z \subset A$.*

It now remains to explain how to handle the probability $\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]$ in the analysis. Notice that from the high probability event $\mathfrak{J}_t$, it suffices to treat the case $\widetilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t \vee \overline{\boldsymbol{\mu}}_t$. We provide here the places where the analysis differs, the rest of the proof remains unchanged.

- We use that $\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t] = \mathbb{P}\left[\forall i \in Z,\ \varepsilon \vee \left(\mu_i^* - \overline{\mu}_{i,t-1}\right) - 0 \vee \left(\theta_{i,t} - \overline{\mu}_{i,t-1}\right) \leq \varepsilon \middle| \mathcal{H}_t\right]$, is a product of functions that are decreasing with respect to $\varepsilon \vee \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)$.
- We use that $\varepsilon \vee \left(\mu_i^* - \overline{\mu}_{i,t-1}\right) \geq g_i^{-1}(u_i \vee g_i(\varepsilon))$ is equivalent to $\mu_i^* - \overline{\mu}_{i,t-1} \geq g_i^{-1}(u_i \vee g_i(\varepsilon))$. Thus, we don't sum on $\mathbf{s}$, and can use Assumption 4 with $\boldsymbol{\lambda} \in \mathbb{R}_+^n$.

Proof of Lemma 8. It is sufficient to prove that

$$\mathfrak{Z}_t, \neg\mathfrak{C}_t \Rightarrow \exists Z \subset A^*,\ Z \neq \emptyset \text{ s.t. } \mathfrak{S}_t(Z) \text{ holds}, \tag{8}$$

because $\neg\mathfrak{C}_t$ and $\mathfrak{S}_t(Z)$ together imply $\mathfrak{T}_t(Z)$. Indeed, see that from $\neg\mathfrak{T}_t(Z)$, we can plug $\boldsymbol{\theta}' = \boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t$ into $\mathfrak{S}_t(Z)$ to get

$$\begin{aligned}
\mathbf{e}_{A_t}^\intercal \widetilde{\boldsymbol{\theta}}_t &= \max_{A \in \mathcal{A}} \mathbf{e}_A^\intercal \widetilde{\boldsymbol{\theta}}_t \\
&\geq \max_{A \in \mathcal{A}} \mathbf{e}_A^\intercal \left(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)
\end{aligned}$$

$$= \mathbf{e}_{\text{Oracle}\left(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)$$
$$> \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon,$$

giving $\mathfrak{C}_t$. To prove (8), we first consider the choice $Z = Z_1 = A^*$. Two cases can be distinguished:

1a) $\forall \boldsymbol{\theta}'$ s.t. $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_{A^*} \leq \varepsilon \mathbf{e}_{A^*}$, we have $A^* \subset A$ for any action $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right)$.

1b) $\exists \boldsymbol{\theta}'$ s.t. $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_{A^*} \leq \varepsilon \mathbf{e}_{A^*}$ such that $A^* \not\subset A$ for some action $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right)$.

1a) For the first case, consider any vector $\boldsymbol{\theta}'$ such that $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_{A^*} \overset{(9)}{\leq} \varepsilon \mathbf{e}_{A^*}$ and let $A \overset{(10)}{=} \text{Oracle}\left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right)$. We can write

$$\mathbf{e}_A^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) \overset{(11)}{\geq} \mathbf{e}_{A^*}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) \overset{(12)}{\geq} \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - m^* \varepsilon,$$

where (11) is from (10), and (12) is from (9). This rewrites as

$$\mathbf{e}_A^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) \geq \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - m^* \varepsilon > \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon,$$

so $\mathfrak{R}_t(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}, A^*)$ holds. Therefore, we have proved that $\mathfrak{S}_t(A^*)$ holds.

1b) For the second case, we have some vector $\boldsymbol{\theta}'$ such that $0 \overset{(13)}{\leq} \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right) \odot \mathbf{e}_{A^*} \overset{(14)}{\leq} \varepsilon \mathbf{e}_{A^*}$, and some action $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right)$ such that $A^* \not\subset A$. We consider $Z_2 = A^* \cap A$. We first prove that $Z_2 \neq \emptyset$ by showing that if an action S' is such that $S' \cap A^* \overset{(15)}{=} \emptyset$, then $A \neq S'$:

$$\begin{aligned}
\mathbf{e}_{S'}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) &\overset{(16)}{=} \mathbf{e}_{S'}^{\mathsf{T}} \widetilde{\boldsymbol{\theta}}_t \overset{(17)}{\leq} \mathbf{e}_{A_t}^{\mathsf{T}} \widetilde{\boldsymbol{\theta}}_t \\
&\overset{(18)}{\leq} \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon \\
&< \mathbf{e}_{A^*}^{\mathsf{T}} \boldsymbol{\mu}^* - m^* \varepsilon \\
&\overset{(19)}{\leq} \mathbf{e}_{A^*}^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right),
\end{aligned}$$

where (16) is from (15), (17) is from the definition of A_t, (18) is from $\neg \mathfrak{C}_t$ and (19) is from (14). Now, we again distinguish two cases:

2a) $\forall \boldsymbol{\theta}''$ s.t. $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}''\right) \odot \mathbf{e}_{Z_2} \leq \varepsilon \mathbf{e}_{Z_2}$, we have $Z_2 \subset B$ for any action $B \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}} \left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right)$.

2b) $\exists \boldsymbol{\theta}''$ s.t. $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}''\right) \odot \mathbf{e}_{Z_2} \leq \varepsilon \mathbf{e}_{Z_2}$ such that $Z_2 \not\subset B$ for some action $B \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}} \left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right)$.

Notice that when $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}''\right) \odot \mathbf{e}_{Z_2} \overset{(20)}{\leq} \varepsilon \mathbf{e}_{Z_2}$, then

$$\mathbf{e}_A^{\mathsf{T}} \left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right) \geq \mathbf{e}_A^{\mathsf{T}} \left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) - (m^* - 1)\varepsilon. \tag{21}$$

Indeed, (21) is a consequence of

$$\mathbf{e}_A^{\mathsf{T}} \left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c} - \boldsymbol{\theta}' \odot \mathbf{e}_{A^*} - \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) = \mathbf{e}_{Z_2}^{\mathsf{T}} \left(\boldsymbol{\theta}'' - \boldsymbol{\theta}'\right)$$

$$= \mathbf{e}_{Z_2}^\mathsf{T}\left(\boldsymbol{\theta}'' - \boldsymbol{\mu}^*\right) + \mathbf{e}_{Z_2}^\mathsf{T}\left(\boldsymbol{\mu}^* - \boldsymbol{\theta}'\right)$$
$$\geq -\varepsilon(m^* - 1) + 0,$$

where we used (20), (13) and that Z_2 is strictly included in A^*.

2a) For the first case, considering any vector $\boldsymbol{\theta}''$ such that $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}''\right) \odot \mathbf{e}_{Z_2} \leq \varepsilon \mathbf{e}_{Z_2}$, we have with $B = \text{Oracle}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right)$ that

$$\begin{aligned}
\mathbf{e}_B^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right) &\geq \mathbf{e}_A^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right) \\
&\overset{(22)}{\geq} \mathbf{e}_A^\mathsf{T}\left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) - (m^* - 1)\varepsilon \\
&\geq \mathbf{e}_{A^*}^\mathsf{T}\left(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\right) - (m^* - 1)\varepsilon \\
&\overset{(23)}{\geq} \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - m^*\varepsilon - (m^* - 1)\varepsilon,
\end{aligned}$$

where (22) uses (21) and (23) uses (14). This rewrites as

$$\mathbf{e}_B^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right) \geq \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - (m^*(m^* + 1)/2 + 1)\varepsilon,$$

so $\mathfrak{R}_t(\boldsymbol{\theta}' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}, Z_2)$ holds, and thus we proved that $\mathfrak{S}_t(Z_2)$ holds.

2b) For the second case, we have a vector $\boldsymbol{\theta}''$ such that $0 \leq \left(\boldsymbol{\mu}^* - \boldsymbol{\theta}''\right) \odot \mathbf{e}_{Z_2} \leq \varepsilon \mathbf{e}_{Z_2}$ and an action $B \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right)$ such that $Z_2 \not\subset B$. We consider $Z_3 = Z_2 \cap B$. Again, $Z_3 \neq \emptyset$ because for any S' such that $S' \cap Z_2 = \emptyset$, we have $S' \neq \text{Oracle}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right)$:

$$\begin{aligned}
\mathbf{e}_{S'}^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right) = \mathbf{e}_{S'}^\mathsf{T}\widetilde{\boldsymbol{\theta}}_t &\leq \mathbf{e}_{A_t}^\mathsf{T}\widetilde{\boldsymbol{\theta}}_t \\
&\leq \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - (m^*(m^* + 1)/2 + 1)\varepsilon \\
&< \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - (m^* + (m^* - 1))\varepsilon \\
&\leq \mathbf{e}_A^\mathsf{T}\left(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\right),
\end{aligned}$$

where the last inequality is obtained in the same way as in inequalities from (22) to (23).

We could repeat the above argument and each time the size Z_i is decreased by at least 1. Thus, after at most $m^* - 1$ steps, since $m^* + (m^* - 1) + (m^* - 2) + \cdots + 1 = m^*(m^* + 1)/2$ is still less than $m^*(m^* + 1)^2/2 + 1$, we could reach the end and find a $Z_i \neq \emptyset$ such that $\mathfrak{S}_t(Z_i)$ holds. □

Proof of Lemma 9. Let's prove that $\arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z) \subset \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\boldsymbol{\eta}$. Consider any action $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z)$. If $A \notin \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\boldsymbol{\eta}$, then there exists $B \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\boldsymbol{\eta}$ such that

$$\mathbf{e}_A^\mathsf{T}\boldsymbol{\eta} < \mathbf{e}_B^\mathsf{T}\boldsymbol{\eta}.$$

Furthermore, since $Z \subset B$ and $\boldsymbol{\delta} \geq 0$, we also have

$$\mathbf{e}_A^\mathsf{T}(\boldsymbol{\delta} \odot \mathbf{e}_Z) \leq \mathbf{e}_B^\mathsf{T}(\boldsymbol{\delta} \odot \mathbf{e}_Z),$$

so we finally have

$$\mathbf{e}_A^\mathsf{T}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z) < \mathbf{e}_B^\mathsf{T}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z),$$

contradicting that $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z)$. □

E General CMAB results

In this section, we state general results that are useful for every regret analysis that we conducted in this paper. The main result of the section is the following theorem, inspired from the analysis of Degenne and Perchet [2016b], that gives a regret bound under the event that the gap Δ_t is controlled by a ℓ_2 norm type error.

Theorem 4 (Regret bound for ℓ_2-norm error)**.** *For all $i \in [n]$, let $\beta_{i,T} \in \mathbb{R}_+$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \left\| \sum_{i \in A_t} \frac{\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2 \right\}.$$

Then,

$$\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\} \Delta_t \leq 32 \log_2^2(4\sqrt{m}) \sum_{i \in [n]} \beta_{i,T} \Delta_{i,\min}^{-1}.$$

Proof. Let $t \geq 1$. We define $\Lambda_t \triangleq \left\| \sum_{i \in A_t} \beta_{i,T}^{1/2} N_{i,t-1}^{-1/2} \mathbf{e}_i \right\|_2$. We start by a simple lower bound on Λ_t, holding for any $j \in A_t$,

$$\Lambda_t \geq \left\| \frac{\beta_{j,T}^{1/2} \mathbf{e}_j}{N_{j,t}^{1/2}} \right\|_2 = \frac{\beta_{j,T}^{1/2}}{N_{j,t}^{1/2}}. \tag{24}$$

We then use the same reverse amortisation technique than in Wang and Chen [2017].

$$\begin{aligned}
\Lambda_t &= -\Lambda_t + \left\| \sum_{i \in A_t} \frac{2\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2 \\
&= -\left\| \frac{\Lambda_t \mathbf{e}_{A_t}}{\|\mathbf{e}_{A_t}\|_2} \right\|_2 + \left\| \sum_{i \in A_t} \frac{2\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2 \\
&\leq \left\| \sum_{i \in A_t} \left(\frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} - \frac{\Lambda_t}{\|\mathbf{e}_{A_t}\|_2} \right)^+ \mathbf{e}_i \right\|_2 \\
&= \left\| \sum_{i \in A_t} \left(\frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} - \frac{\Lambda_t}{\|\mathbf{e}_{A_t}\|_2} \right)^+ \mathbb{I}\left\{ \Lambda_t \geq \frac{\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} \right\} \mathbf{e}_i \right\|_2 && \text{Using (24)} \\
&\leq \left\| \sum_{i \in A_t} \mathbb{I}\left\{ 2\Lambda_t \geq \frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} \geq \frac{\Lambda_t}{\|\mathbf{e}_{A_t}\|_2} \right\} \frac{2\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2.
\end{aligned}$$

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

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

This induces a partition of the set of indices:

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

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

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

We can thus upper bound Λ_t^2 using this decomposition

$$\Lambda_t^2 \leq \left\| \sum_{i \in A_t} \mathbb{I}\left\{ 2\Lambda_t \geq \frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} \geq \frac{\Lambda_t}{\|\mathbf{e}_{A_t}\|_2} \right\} \frac{2\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2^2$$

$$\leq \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2)\rceil} \left\| \sum_{i\in J_{k,t}} \frac{2\beta_{i,T}^{1/2}\mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2^2$$
$$\leq \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2)\rceil} 2^{2-2k}\Lambda_t^2 \left\|\mathbf{e}_{J_{k,t}}\right\|_2^2.$$

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

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

$$\begin{aligned}
\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t &\leq \sum_{t=1}^T \sum_{k=0}^{\lceil\log_2(m)/2\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\}\Delta_t \\
&\leq \sum_{t=1}^T \sum_{k=0}^{\lceil\log_2(m)/2\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\} \sum_{i\in I} \mathbb{I}\{i\in J_{k,t}\}\Delta_t 2^{2-2k}(\lceil\log_2(m)/2\rceil + 1) \\
&\leq \sum_{t=1}^T \sum_{k=0}^{\lceil\log_2(m)/2\rceil} \sum_{i\in I} \mathbb{I}\left\{i\in A_t,\ N_{i,t-1}^{1/2} \leq \frac{2^{k+1}\beta_{i,T}^{1/2}}{\Delta_t}\right\}\Delta_t 2^{2-2k}(\lceil\log_2(m)/2\rceil + 1) \\
&= (\lceil\log_2(m)/2\rceil + 1) \sum_{k=0}^{\lceil\log_2(m)/2\rceil} 2^{2-2k} \sum_{i\in I} \underbrace{\sum_{t=1}^T \mathbb{I}\left\{i\in A_t,\ N_{i,t-1}^{1/2} \leq \frac{2^{k+1}\beta_{i,T}^{1/2}}{\Delta_t}\right\}\Delta_t}_{(25)_{i,k}}.
\end{aligned}$$

Applying Proposition 2 gives

$$(25)_{i,k} \leq \frac{\beta_{i,T} 2^{\frac{k+1}{1/2}}}{1-1/2}\Delta_{i,\min}^{1-1/1/2}.$$

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

$$\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t \leq 32\log_2^2(4\sqrt{m}) \sum_{i\in[n]} \beta_{i,T}\Delta_{i,\min}^{-1}.$$

□

The following Proposition 2 is a standard and general result in CMAB, that was first proved in Chen et al. [2013].

Proposition 2. *Let $i \in [n]$ and $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing function, integrable on $[\Delta_{i,\min}, \Delta_{i,\max}]$. Then*

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

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 $i \in A_t$. We define a dummy gap $\Delta_{i,0} = \infty$ and let $f_i(\Delta_{i,0}) = 0$. In (26), 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 $i \in A_t$. Notice that in (26), 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 $k \in [k_t]$.

$$\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ N_{i,t-1} \leq f_i(\Delta_t)\}\Delta_t$$
$$= \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k}), \Delta_{i,k} \geq \Delta_t\}\Delta_t. \tag{26}$$

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 $\Delta_{i,k}$.

$$(26) \leq \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k}), \Delta_{i,k} \geq \Delta_t\}\Delta_{i,k}. \tag{27}$$

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 $\{\Delta_{i,k} \geq \Delta_t\}$.

$$(27) \leq \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k})\}\Delta_{i,k}. \tag{28}$$

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

$$(28) = \sum_{k=1}^{K_i}\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ f_i(\Delta_{i,k-1}) < N_{i,t-1} \leq f_i(\Delta_{i,k})\}\Delta_{i,k}. \tag{29}$$

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 $\{i \in A_t\}$, 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 $i \in A_t$. 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 $\lfloor y \rfloor - \lfloor x \rfloor$.

$$(29) \leq \sum_{k=1}^{K_i} (\lfloor f_i(\Delta_{i,k}) \rfloor - \lfloor f_i(\Delta_{i,k-1}) \rfloor)\Delta_{i,k}. \tag{30}$$

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

$$(30) = \lfloor f_i(\Delta_{i,K_i}) \rfloor \Delta_{i,K_i} + \sum_{k=1}^{K_i-1} \lfloor f_i(\Delta_{i,k}) \rfloor (\Delta_{i,k} - \Delta_{i,k+1}) \tag{31}$$

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 $k \in [K_i - 1]$.

$$(31) \leq f_i(\Delta_{i,K_i})\Delta_{i,K_i} + \sum_{k=1}^{K_i-1} f_i(\Delta_{i,k})(\Delta_{i,k} - \Delta_{i,k+1}) \tag{32}$$
$$\leq f_i(\Delta_{i,K_i})\Delta_{i,K_i} + \int_{\Delta_{i,K_i}}^{\Delta_{i,1}} f_i(x)\mathrm{d}x. \tag{33}$$

□

There also exist a version for the ℓ_1-norm error.

Theorem 5 (Regret bound for ℓ_1-norm error)**.** *For all $i \in [n]$, let $\beta_{i,T} \in \mathbb{R}_+$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \left\| \sum_{i \in A_t} \frac{\beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_1 \right\}.$$

Then,

$$\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\} \Delta_t \leq \sum_{i \in [n]} \beta_{i,T} 8m {\Delta_{i,\min}}^{-1}.$$

Proof. Let $t \geq 1$. The first step is the reverse amortisation technique, that allows us to modify the upper bound on Δ_t in such a way that indices i such that $N_{i,t-1}$ is high enough are removed. Assuming that $\mathfrak{A}_t$ holds, we get

$$\Delta_t \leq \sum_{i \in A_t} \mathbb{I}\left\{ \frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}} \geq \frac{\Delta_t}{m} \right\} \frac{2\beta_{i,T}^{1/2}}{N_{i,t-1}^{1/2}}$$

Now, we apply Proposition 3. In summary, we have that $\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\} \Delta_t$ is upper bounded by

$$\sum_{i \in [n]} \beta_{i,T} 8m {\Delta_{i,\min}}^{-1}.$$

□

Proposition 3. *Let $i \in [n]$ and $f_i(x) = \beta_{i,T} x^{-1/\alpha_i}$, $\alpha_i \in (0,1]$ and $\beta_{i,T} \geq 0$. Then*

$$\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ \delta_t \neq 0,\ N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1}) \leq \delta_{i,\min}^{1-1/\alpha_i} \frac{\beta_{i,T}}{1-\alpha_i} \mathbb{I}\{\alpha_i < 1\} + \mathbb{I}\{\alpha_i = 1\} \beta_{i,T} \left(1 + \log\left(\frac{\beta_{i,T}}{\delta_{i,\min}}\right)\right).$$

Proof. We upper bound $f_i(\delta_t)$ by $f_i(\delta_{i,\min})$ directly in the event, and then simply count the number of integers in $(0, f_i(\delta_{i,\min})]$. For each such integer s, the regret suffered is $f_i^{-1}(s)$. We then upper bound the sum by an integral (using the fact that f_i^{-1} is decreasing), to get the final result.

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ \delta_t \neq 0,\ N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1}) &\leq \sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ N_{i,t-1} \leq f_i(\delta_{i,\min})\} f_i^{-1}(N_{i,t-1}) \\
&\leq \sum_{s=1}^{\lfloor f_i(\delta_{i,\min}) \rfloor} f_i^{-1}(s) \\
&\leq f_i^{-1}(1) + \int_1^{f_i(\delta_{i,\min})} f_i^{-1}(s) \mathrm{d}s \\
&= \beta_{i,T}^{\alpha_i} + \int_1^{\beta_{i,T} \delta_{i,\min}^{-1/\alpha_i}} \beta_{i,T}^{\alpha_i} s^{-\alpha_i} \mathrm{d}s \\
&\leq \mathbb{I}\{\alpha_i < 1\} \delta_{i,\min}^{1-1/\alpha_i} \frac{\beta_{i,T}}{1-\alpha_i} \\
&\quad + \mathbb{I}\{\alpha_i = 1\} \beta_{i,T} \left(1 + \log\left(\frac{\beta_{i,T}}{\delta_{i,\min}}\right)\right).
\end{aligned}$$

□

Proposition 4 (Regret bound for a composed bonus)**.** *Let $K \in \mathbb{N}^*$. For all $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \sum_{k \in [K]} B_{k,t} \right\},$$

for some $B_{k,t} \geq 0$. Then, the event-filtered regret $\mathbb{E}\left[\sum_{t=1}^{T} \Delta_t \mathbb{I}\{\mathfrak{A}_t\}\right]$ is upper bounded by

$$\sum_{k \in [K]} \mathbb{E}\left[\sum_{t \in [T]} \Delta_t \mathbb{I}\{\Delta_t \leq K B_{k,t}\} \right].$$

Proof. From $\mathfrak{A}_t$, there must exists one k such that $\Delta_t \leq K B_{k,t}$. So $1 \leq \sum_{k \in [K]} \mathbb{I}\{\Delta_t \leq K B_{k,t}\}$, i.e., $\Delta_t \leq \sum_{k \in [K]} \Delta_t \mathbb{I}\{\Delta_t \leq K B_{k,t}\}$. □