Efficient Learning in Stochastic Combinatorial Semi-Bandits

Pierre Perrault

université
PARIS-SACLAY

Apprentissage Efficient dans les Problèmes de Semi-Bandits Stochastiques Combinatoires

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques appliquées

Unité de recherche : Centre Borelli (ENS Paris-Saclay), UMR 9010 CNRS
Centre Borelli, 91190, Gif-sur-Yvette, France
Référent : Ecole normale supérieure de Paris-Saclay

Thèse présentée et soutenue en visioconférence, le 30/11/2020, par

Pierre PERRAULT

Composition du jury :

Richard Combes Professeur Assistant, Centrale Supelec L2S	Examinateur
Raphaël Féraud Chercheur, Orange Labs	Invité
Marc Lelarge Directeur de recherche, INRIA	Rapporteur, Président du jury
Vianney Perchet Professeur, CREST, ENSAE	Codirecteur
Alexandre Proutiere Professeur, KTH EECS, Stockholm	Rapporteur
Michal Valko Chercheur, DeepMind	Directeur
Claire Vernade Attachée de recherche, DeepMind	Examinatrice

Thèse de doctorat

NNT: 2020UPASM023


école
normale
supérieure
paris—saclay
Fondation mathématique
FMJH
Jacques Hadamard
Inria
CRIStAL

UNIVERSITÉ PARIS-SACLAY

Abstract

INRIA Lille - Nord Europe
Sequential Learning (SequeL) team

Efficient Learning in Stochastic Combinatorial Semi-Bandits

by Pierre PERRAULT

Combinatorial stochastic semi-bandits appear naturally in many contexts where the exploration/exploitation dilemma arises, such as web content optimization (recommendation/online advertising) or shortest path routing methods. This problem is formulated as follows: an agent sequentially optimizes an unknown and noisy objective function, defined on a power set $\mathcal{P}([n])$. For each set A tried out, the agent suffers a loss equal to the expected deviation from the optimal solution while obtaining observations to reduce its uncertainty on the coordinates from A. Our objective is to study the efficiency of policies for this problem, focusing in particular on the following two aspects: statistical efficiency, where the criterion considered is the regret suffered by the policy (the cumulative loss) that measures learning performance; and computational efficiency. It is sometimes difficult to combine these two aspects in a single policy. In this thesis, we propose different directions for improving statistical efficiency, while trying to maintain the computational efficiency of policies. In particular, we have improved optimistic methods by developing approximation algorithms and refining the confidence regions used. We also explored an alternative to the optimistic methods, namely randomized methods, and found them to be a serious candidate for combining the two types of efficiency.

Acknowledgements - Remerciements

Avant tout, je voudrais exprimer ma profonde et sincère gratitude à Michal et Vianney, pour leur accompagnement et leur support pendant ces trois années. Travailler avec vous a été un réel plaisir. Vous avez su me faire découvrir le monde de la recherche avec beaucoup de bienveillance et d'engagement, le tout dans une ambiance de travail conviviale et ludique. Je me suis vraiment senti libre concernant le choix des directions de recherche, et vous m'avez à chaque fois encouragé à privilégier celles que j'aimais, sans rien imposer, mais toujours avec de précieux conseils. Je vous en suis très reconnaissant car je pense que c'est l'ingrédient essentiel d'une thèse épanouissante.

Michal, I especially wanted to thank you for your investment in the success of my thesis. You have always opened me to the best opportunities, for example by allowing me to co-teach your *Graphs in Machine Learning* course, or by helping me find my internship at Adobe in San-José! You also have contributed a lot to my social life outside of research, I am thinking of the squash, swimming pool and beer sessions (by the way, thanks again for the racquet you offered me when my first paper got accepted, we absolutely have to test it together after the Covid!).

Vianney, ton côté très détendu en cherchant toujours à s'attaquer aux problèmes les plus "marrants" me restera longtemps en mémoire. Je pense en particulier aux longues discussions avec Emmanuel et Thomas devant ton écran tactile géant, à essayer de caractériser l'ordre optimal dans le "problème du secrétaire" (je pense également à Mathilde, qui a été à l'origine de ces discussions de mon point de vue), ou encore à mon stage du MVA, où le but était de "trouver des gens dans des arbres". J'approuve totalement cette façon de considérer la recherche comme un jeu, ce qui n'empêche pas d'obtenir des résultats intéressants !

Je tiens aussi à remercier tout particulièrement Marc, Alexandre, Claire, Richard et Raphaël d'avoir accepté d'évaluer ma thèse. Ce fut un honneur et un privilège de vous avoir en tant que membres du jury, tant vos travaux sur les bandits ont grandement inspiré mes recherches. J'ai également beaucoup apprécié vos nombreux retours, ainsi que l'intérêt que vous avez porté à mon manuscrit et à ma présentation.

Jennifer and Zheng, thank you so much for hosting me and teaching me a lot during my internship at Adobe! It was a very rewarding experience for me, both on a scientific level, with our double research project, and on a human level, with our trips to San-Francisco or our rodeo on the mechanical bull with Georgios! Hope to see you all very soon!

Je voulais également remercier tous les chercheurs et doctorants de l'équipe Sequel/Scool et de l'équipe du CMLA/Centre Borelli. J'ai apprécié nos discussions de groupe, surtout lorsqu'il s'agissait de questions mathématiques épineuses (dédicace à Valentin, qui maîtrise vraiment cet art !). Merci à toi Étienne pour la qualité de tes réflexions, tant lors de notre collaboration que pendant les "group meetings". Mes pensées vont aussi à Xavier, mon "binôme" de thèse. J'ai vraiment apprécié notre coopération. Je me souviendrai longtemps des bons moments passés au quotidien avec toi, Tina, Jérémie, Mariano et Roger, dans notre bonne vieille "salle des doctorants", où l'ambiance n'était pas toujours des plus sérieuses[1] !

Je veux aussi saluer ceux qui me sont chers, à savoir ma famille et mes amis. Leurs attentions et leurs encouragements m'ont accompagné tout le long de cette aventure. Vous m'avez également permis de décompresser et de faire la fête ! Pour tous ces bons moments, je vous dis merci ! Mes pensées vont en particulier à mes parents, mes

[1] Batailles de claque-doigts, tests des fréquences de résonance des salles, jeux de société, les anecdotes ne manquent pas !

grands-parents et mes frères, qui se sont toujours intéressés à mes travaux, et ont fait leur possible pour créer un environnement propice à ma réussite. Enfin, un petit mot pour Julia, ma petite nièce adorée : j'ai hâte de pouvoir t'expliquer ma thèse quand tu seras assez grande !

Par-dessus tout, je tiens à témoigner ma reconnaissance à ma Mégane ! Tu as su être ma principale source d'inspiration et de soutien dans mes recherches. C'est à toi que je veux dédier ce travail, car tu y as largement contribué, notamment en m'apportant chaque jour un peu plus de bonheur.

Contents

Chapter 1

Introduction

1.1 Notations

We present here the general notation used in this thesis. We will also add more specific notations when the corresponding concepts are introduced.

For two real numbers x and y, we use the notation $x \vee y \triangleq \max\{x, y\}$ and $x \wedge y \triangleq \min\{x, y\}$. We also define $\text{sign}(x)$ to be 0 if $x = 0$ and $x/|x|$ otherwise. We denote by $|A|$ the cardinality of a set A, and by A^c the complement of A when the ground set is clear from the context. For any integer $n \in \mathbb{N}^*$, the set of integers between 1 and n is denoted by $[n] \triangleq \{1, \dots, n\}$, and the associated power set is denoted $\mathcal{P}([n]) \triangleq \{A,\ A \subset [n]\}$. We denote the Minkowski sum of two sets $Z, Z' \subset \mathbb{R}^n$ as $Z + Z' \triangleq \{z + z', z \in Z, z' \in Z'\}$, and $z + Z' \triangleq \{z\} + Z'$. We typeset vectors and matrices in bold and indicate components with indices, e.g., for some set I, $\mathbf{a} = (a_i)_{i \in I} \in \mathbb{R}^I$ is a vector on I. When there is no ambiguity on the index set, we simply use the notation (a_i). We denote by $\mathbf{a} \odot \mathbf{b} \triangleq (a_i b_i)$ the Hadamard product of two vectors $\mathbf{a}$ and $\mathbf{b}$. The Hadamard product of two matrices is simply the entrywise product, producing another matrix of the same dimension as the operands. We let $\mathbf{e}_{n,i}$ be the i^{th} canonical unit vector of $\mathbb{R}^n$. The incidence vector of any subset $A \in \mathcal{P}([n])$ is $\mathbf{e}_{n,A} \triangleq \sum_{i \in A} \mathbf{e}_{n,i}$. When the dimension is clear from the context, we omit the n and use the notation $\mathbf{e}_A$ and $\mathbf{e}_i$. If $\mathbf{x}, \mathbf{y}$ are two real-valued vectors, we write $\mathbf{x} \geq \mathbf{y}$ if $\mathbf{x} - \mathbf{y}$ has components in $\mathbb{R}_+$, and use $\mathbf{x} \vee \mathbf{y} \triangleq (x_i \vee y_i)_i$ (resp. $\mathbf{x} \wedge \mathbf{y} \triangleq (x_i \wedge y_i)_i$). For a vector $\mathbf{a}$, and $p \geq 1$, we define the ℓ_p-norm of $\mathbf{a}$ as $\|\mathbf{a}\|_p \triangleq (\sum_i |a_i|^p)^{1/p}$, and the ℓ_∞-norm of $\mathbf{a}$ as $\|\mathbf{a}\|_\infty \triangleq \sup_i |a_i|$. We also define $\|\mathbf{a}\|_0 \triangleq |\{i,\ a_i \neq 0\}|$. We denote by $\text{diag}(\mathbf{a})$ the diagonal matrix with the elements $\mathbf{a} = (a_i)$ on the diagonal. We write $\mathbf{I}_n \triangleq \text{diag}\Big(\mathbf{e}_{n,[n]}\Big)$, and only $\mathbf{I}$ when the dimension is clear from the context. We use $\mathbf{A} \succeq \mathbf{B}$ to indicate the Löwner ordering of two $n \times n$ matrices (Horn and Johnson, 1990), i.e., that $\mathbf{A} - \mathbf{B}$ is a positive semi-definite (PSD) matrix. Furthermore, $\mathbf{A} \succeq_+ \mathbf{B}$ means $\boldsymbol{\lambda}^\intercal(\mathbf{A} - \mathbf{B})\boldsymbol{\lambda} \geq 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n_+$, and $\mathbf{A} \succeq_A \mathbf{B}$ (resp. $\mathbf{A} \succeq_{+A} \mathbf{B}$) means $\boldsymbol{\lambda}^\intercal(\mathbf{A} - \mathbf{B})\boldsymbol{\lambda} \geq 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n$ (resp. $\boldsymbol{\lambda} \in \mathbb{R}^n_+$), with $\lambda_i = 0$ for $i \notin A$. Obviously, we also use the notations $\mathbf{y} \leq \mathbf{x}$, $\mathbf{B} \preceq, \preceq_+, \preceq_A, \preceq_{+A} \mathbf{A}$. For any $n \times n$ matrix $\mathbf{A} \succeq 0$, we define the euclidean seminorm associated to $\mathbf{A}$ as $\|\mathbf{x}\|_{\mathbf{A}} \triangleq \sqrt{\mathbf{x}^\intercal \mathbf{A} \mathbf{x}}$, where $\mathbf{x}$ is a vector of $\mathbb{R}^n$.

Events are indicated using the fraktur font $\mathfrak{A}$. $\mathbb{I}\{\mathfrak{A}\}$ is the $\{0, 1\}$ indicator of the event $\mathfrak{A}$, i.e.,

$$\mathbb{I}\{\mathfrak{A}\} = \begin{cases} 1 & \text{if } \mathfrak{A} \text{ holds} \\ 0 & \text{otherwise.} \end{cases}$$

We denote by $\neg\mathfrak{A}$ the complementary event of $\mathfrak{A}$. The following definitions are stated for a random variable $X \in \mathbb{R}$ but can also be extended for a random vector $\mathbf{X} \in \mathbb{R}^n$. We denote by $\mathbb{P}_X$ the probability distribution of X. For another random variable Y, we write $Y \sim \mathbb{P}_X$ (or $Y \sim X$) to indicate that $\mathbb{P}_Y = \mathbb{P}_X$. For a sequence of random

variables (X_t), $\sigma(X_1, \dots, X_t)$ is the sigma-algebra generated by random variables $X_1, \dots, X_t$. We write $(X_t) \sim \mathbb{P}_X$ to indicate that $\mathbb{P}_{X_t} = \mathbb{P}_X$ for all $t \in \mathbb{N}^*$, and $(X_t) \overset{iid}{\sim} \mathbb{P}_X$ to indicate that $\mathbb{P}_{(X_1, X_2, \dots, X_t)} = \mathbb{P}_X^{\otimes t}$ for all $t \in \mathbb{N}^*$. For two familly of distributions $\mathcal{D}_1, \mathcal{D}_2$, we use the notation $\mathcal{D}_1 \otimes \mathcal{D}_2 \triangleq \{P_1 \otimes P_2,\ P_1 \in \mathcal{D}_1 \text{ and } P_2 \in \mathcal{D}_2\}$.

For any finite set $\mathcal{X}$, and any $\boldsymbol{p} \in [0,1]^{\mathcal{X}}$, such that $\sum_{x \in \mathcal{X}} p_x = 1$, the distribution $\sum_{x \in \mathcal{X}} p_x \delta_x$ is a discrete probability measure on $\mathcal{X}$. For example, δ_x is the Dirac distribution at x, and for some $p \in [0,1]$, Bernoulli$(p) = p\delta_1 + (1-p)\delta_0$. For $\mathbf{x} \in [0,1]^n$ with $\sum_i x_i \leq 1$, we define the multinoulli distribution (also called categorical distribution) as Multinoulli$(\mathbf{x}) = \sum_i x_i \delta_{\mathbf{e}_i} + (1 - \sum_i x_i)\delta_{0_{\mathbb{R}^n}}$.

1.2 Présentation du contenu de la thèse (en Français)

Cette thèse étudie les problèmes d'optimisation combinatoire séquentielle dans un environnement stochastique, avec des observations que l'on qualifie de "semi-bandits". La terminologie "semi-bandits" fait reférence au fameux problème du bandit manchot stochastique (aussi appelé bandit stochastique à plusieurs bras). Les problèmes de bandit forment une classe importante de problèmes d'optimisation séquentielle, et ils ont été largement étudiés dans le domaine des statistiques et de l'apprentissage automatique. A l'origine, ils ont été introduits dans l'article fondateur de Robbins (1952) pour étudier les plans d'expériences séquentielles. La version classique du problème est formulée comme un système de n bras (ou machines). Un *agent* doit sélectionner à plusieurs reprises un bras parmi les n disponibles. Chaque bras i a une loi de probabilité inconnue $\mathbb{P}_{X_i}$, de moyenne inconnue μ_i^*, où la variable X_i encode une *récompense*. Après avoir sélectionné un bras, l'agent observe une réalisation indépendante de la distribution de récompense correspondante. Chaque décision est basée sur les décisions passées et les récompenses observées. La tâche pour l'agent consiste à jouer tour à tour les bras de manière à maximiser l'espérance de la récompense cumulée. L'agent doit jouer en équilibrant l'exploitation et l'exploration. En effet, les bras dont les récompenses observées sont les plus élevées doivent être sélectionnés souvent, tandis que tous les bras doivent être explorés pour connaître leurs récompenses moyennes. De manière équivalente, la performance de l'agent (ou de la politique) peut être évaluée par son *regret* R_T, définie comme l'espérance de l'écart sur T tours ($T \in \mathbb{N}^*$ étant appelé l'horizon) entre la récompense accumulée par la politique et celle accumulée par une politique *oracle* sélectionnant toujours le meilleur bras. Le regret peut se réécrire comme

$$R_T \triangleq \sum_{i \in [n]} \mathbb{E}[N_{i,T}]\Delta_i,$$

où $\Delta_i \triangleq \max_j \mu_j^* - \mu_i^*$ est l'écart des moyennes entre le bras optimal et le bras i, et où $N_{i,t}$ est le nombre de fois où le bras i a été tiré jusqu'à l'instant $t \in \mathbb{N}^*$. La notion de regret quantifie ainsi la perte due à la nécessité d'apprendre les récompenses moyennes des différents bras.

Le problème défini ci-dessus est un exemple de problème d'apprentissage par renforcement. En effet, un problème de bandit manchot peut être vu comme un processus de décision markovien avec un seul état. Mentionnons au passage que le nom du problème vient du fait d'imaginer un joueur face à une rangée de machines à sous, devant décider quelles machines jouer, combien de fois jouer à chaque machine et dans quel ordre les jouer. Bien entendu, il ne s'agit que d'une modélisation, et ce problème aborde en pratique tout type de situation où un agent tente simultanément

d'acquérir de nouvelles connaissances (exploration) et d'optimiser ses décisions sur la base des connaissances existantes (exploitation). Il existe de nombreuses applications pratiques du modèle du bandit (manchot), par exemple :

- Les essais cliniques qui étudient les effets de différents traitements expérimentaux tout en minimisant les vies humaines perdues (Thompson, 1933; Gittins, Weber, et Glazebrook, 1989; Berry et Fristedt, 1985).
- La conception de portefeuilles financiers (Hoffman, Brochu, et Freitas, 2011).
- La prise de décision pour l'accès dynamique au spectre dans le contexte des radios cognitives (Lai, Jiang, et Poor, 2008).

Le problème des bandits a notamment été étudié par Lai et Robbins (1985). Ils ont prouvé une borne inférieure asymptotique sur le regret R_T. Cette borne indique que toute politique "raisonnable" (en un certain sens) doit subir au moins un regret logarithmique par rapport à l'horizon T. Elle fournit ainsi une limite de performance fondamentale qu'aucune politique "raisonnable" ne peut battre. La constante dans la borne inférieure est linéaire en n, le nombre de bras. Plus précisement, elle vaut la somme sur les bras sous-optimaux i de l'écart Δ_i fois l'inverse de la divergence de Kullback-Leibler (KL) (Kullback et Leibler, 1951) entre la distribution du bras i et celle du bras optimal. Grossièrement, $\log(T)$ fois l'inverse de la KL représente le nombre de tours nécessaires pour distinguer le bras i du bras optimal, donnant ainsi une interprétation claire de la borne inférieure. Dans le cas de distributions Gaussiennes, l'inverse de la KL est minoré (à une constante multiplicative près) par la variance σ_i^2 du bras i divisée par Δ_i^2. En simplifiant, la borne inférieure devient donc

$$\Omega\left(\log(T) \sum_{i\in[n],\ \Delta_i>0} \frac{\sigma_i^2}{\Delta_i}\right). \tag{1.1}$$

Il n'est pas surprenant que la variance apparaisse, car elle peut être considérée comme une mesure de l'incertitude des échantillons : plus la variance est élevée, plus l'estimation est difficile, et donc plus le regret doit être élevé. Lai et Robbins (1985) ont également élaboré des politiques pour certaines distributions de récompense et ont montré qu'elles atteignent asymptotiquement la borne inférieure. En effet, ces politiques ont une borne supérieure sur leur regret qui est logarithmique en T, avec une constante du même ordre que dans la borne inférieure. Dans le cas où les bras ont des distributions σ_i^2-*sous-Gaussiennes*, la stratégie UCB (pour *Upper Confidence Bound* (Auer, Cesa-Bianchi, et Fischer, 2002)) atteint asymptotiquement la borne (1.1). Elle consiste à choisir, au tour t, le bras i_t maximisant la borne supérieure de l'intervalle de confiance $\overline{\mu}_{i,t-1} + \sqrt{2\log(t)/N_{i,t-1}}$, où $\overline{\mu}_{i,t-1}$ est la moyenne empirique des récompenses reçues en ayant tiré le bras i. Il s'agit d'une stratégie dite *optimiste dans l'incertain*, car elle choisit le bras qui serait le meilleur si les moyennes vallaient leurs estimations optimistes.

Bandit stochastique combinatoire Pour revenir au sujet de cette thèse, à savoir les problèmes d'optimisation séquentielle combinatoire, nous pouvons étendre le cadre mentionné ci-dessus : l'agent peut désormais, lors d'un tour, choisir plusieurs bras au lieu d'un seul (avec éventuellement quelques contraintes sur les ensembles de bras possibles). La récompense obtenue est fonction de la réalisation individuelle de chaque bras choisi. Ceci est bien adapté aux applications où les actions dont dispose l'agent

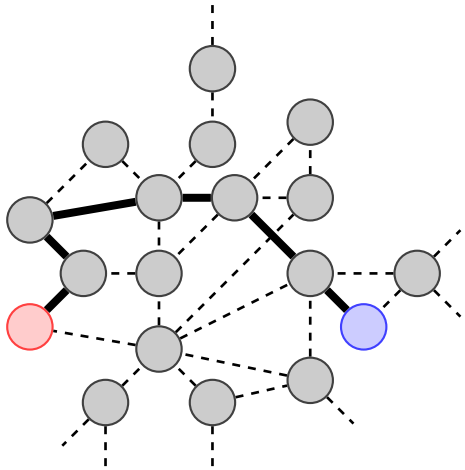

FIGURE 1.1: Exemple d'un réseau (graphe) où chaque arête est un bras. Le nœud en bleu représente l'origine, et le nœud en rouge la destination. Les arêtes représentées par des lignes en gras forment un chemin entre l'origine et la destination : il s'agit donc d'une action que l'agent peut décider de jouer lors d'un tour. Se faisant, chaque bras du chemin produit une réalisation, et la récompense peut alors être la somme des réalisations du chemin.

appartiennent à un espace combinatoire et sont construites à partir de plusieurs petites actions (les bras). Mentionnons par exemple le problème du plus court chemin entre deux points d'un réseau (qui a été étudié notamment par Liu et Zhao (2012), Talebi, Zou, et al. (2013)). Chaque arête du réseau est un bras auquel est associé une loi de probabilité, modélisant par exemple la fluidité du trafic sur cette arête. L'agent choisit à chaque tour un chemin reliant les deux points, et obtient en récompense une réalisation indépendante de la fluidité globale du chemin choisi. Cette fluidité globale est définie selon le contexte, et peut être par exemple la somme des fluidités des arêtes du chemin (voir Figure 1.1). Ce cas particulier où la récompense est la somme des réalisations des bras choisis correspond à un problème d'optimisation combinatoire séquentiel avec une fonction objectif *linéaire*, et il s'agit du scénario le plus répandu parmi ces types de problèmes. Il est étudié, par exemple, par Abernethy, Hazan, et Rakhlin (2008), Dani, Hayes, et Kakade (2008), ou encore Bubeck, Cesa-Bianchi, et Kakade (2012). De manière générale, l'extension combinatoire du problème de bandit est connue dans la littérature sous le nom de *bandit stochastique combinatoire*. Elle implique (au moins) deux cadres distincts concernant les observations dont dispose l'agent sur les réalisations des bras à chaque tour :

- *Observation "semi-bandit"* : toutes les réalisations des bras choisis sont observées par l'agent.
- *Observation "bandit"* : seule la récompense (qui est fonction des réalisations des bras choisis) est observée.

Comme nous l'avons déjà mentionné, nous nous intéressons ici au premier cadre, communément appelé *semi-bandits stochastiques combinatoires* (que nous abrégerons en CMAB dans cette thèse, pour *combinatorial multi-armed bandits*, le "semi" étant omis par souci de concision). Puisque l'agent a plus d'observations, cette hypothèse peut être plus difficile à satisfaire dans la pratique, mais peut conduire à des politiques plus performantes. Notons aussi que le second cadre est en fait un cas particulier du

problème de bandit introduit plus haut, où le nombre de bras est certes combinatoire, mais où une certaine structure est spécifiée entre les récompenses.

1.2.1 Observation "semi-bandit"

Nous formalisons ici le problème CMAB. La réalisation du bras $i \in [n]$ au tour t est notée $X_{i,t} \in \mathbb{R}$. Les vecteurs aléatoires $(\mathbf{X}_t)_{t\geq 1}$ de $\mathbb{R}^n$ sont i.i.d., avec une distribution inconnue $\mathbb{P}_{\mathbf{X}}$, de moyenne $\boldsymbol{\mu}^* \in \mathbb{R}^n$. Néanmoins, sauf mention contraire, X_i et X_j pour deux bras distincts $i \neq j$ peuvent être arbitrairement corrélées. L'action (aussi appelée *super-bras*) sélectionnée par l'agent au tour t (i.e., l'ensemble des bras choisis) est notée A_t. L'ensemble des super-bras possibles est appelé *espace d'action*, et est noté $\mathcal{A}$. C'est un sous-ensemble fixé de $\mathcal{P}([n])$, tel que chacun de ses éléments A est un sous-ensemble d'au plus m bras. Après avoir sélectionné un bras A_t au tour t, l'agent reçoit comme retour d'information $\mathbf{e}_{A_t} \odot \mathbf{X}_t$. Bien que nous considérons un cadre plus général dans la thèse, afin de simplifier les explications, nous nous limitons ici au cas d'une fonction de récompense linéaire : la récompense au tour t est donc $\mathbf{e}_{A_t}^\intercal \mathbf{X}_t$. Nous supposons aussi que $\mathcal{A}$ est tel que l'optimisation de fonctions linéaires peut être fait de manière efficace (en temps polynomial en n).

L'objectif est d'identifier une politique qui maximise la récompense cumulée espérée sur l'ensemble des T tours. L'espérance est prise sur l'aléa des récompenses et sur l'éventuel aléa de la politique suivie par l'agent (une action peut être choisie au hasard). De manière équivalente, nous visons à concevoir une politique π qui minimise le regret, défini par :

$$R_T(\pi) \triangleq \max_{A\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} (\mathbf{e}_A - \mathbf{e}_{A_t})^\intercal \mathbf{X}_t\right].$$

Contrairement au problème de bandit classique, nous pouvons remarquer qu'il y a une notion de similarité entre les actions. En effet, deux actions "proches", c'est-à-dire ayant beaucoup de bras en commun, ont tendance à avoir une récompense "proche". Cette structure peut être exploitée dans l'apprentissage, afin de contrebalancer la taille combinatoire de l'espace d'action. Il en résulte la conception de politiques efficaces de sélection des super-bras, basée sur le principe d'optimisme dans l'incertain, dont le regret est de l'ordre de $\mathcal{O}(n \cdot m \log(T)/\Delta)$ (Kveton, Wen, Ashkan, and Szepesvari, 2015b), où $\Delta > 0$ est la différence minimale entre la moyenne d'un super-bras optimal et la moyenne d'un super-bras sous-optimal. Ces politiques atteignent asymptotiquement une borne inférieure : en considérant l'espace d'action $\mathcal{A} = \{A_1, \ldots, A_{n/m}\}$ avec $A_k = \{(k-1)m+1, \ldots, km\}$, et en posant pour tout k, $X_{(k-1)m+1} = \cdots = X_{km}$, tirée selon une Gaussienne de variance 1, on réduit le problème à un problème de bandit classique, permettant d'appliquer la borne (1.1), avec $\sigma_i^2 = m^2$. La politique la plus connue de ce type est CUCB (*Combinatorial Upper Confidence Bound*), et est fortement inspirée de UCB. En effet, elle joue au tour t

$$A_t \in \arg\max_{A\in\mathcal{A}} \mathbf{e}_A^\intercal \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{2\log(t)}{N_{i,t-1}}}\right)_i,$$

en d'autres termes, les estimations optimistes sont les mêmes que pour UCB. Notons que dans les problèmes CMAB, la distribution conjointe tout entière du vecteur des réalisations $\mathbf{X}$ a de l'importance, contrairement aux bandits classiques où seulement les marginales sont suffisantes pour caractériser une instance d'un problème. Ainsi, lorsqu'il est question d'estimer le vecteur des moyennes $\boldsymbol{\mu}^*$, nous sommes davantage

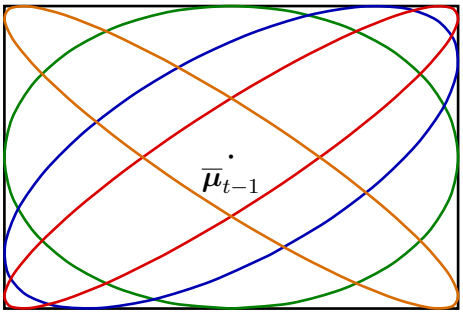

FIGURE 1.2: La région de confiance de type hypercube (en noir) correspond à un a priori Gaussien pour chaque marginale. La corrélation entre ces marginales peut cependant être arbitraire : voici plusieurs exemples, de très négativement corrélées (en orange) à très positivement corrélées (en rouge), en passant par indépendantes (en vert). On voit donc que tout l'hypercube est susceptible de contenir la vraie moyenne $\boldsymbol{\mu}^*$ tant que l'on ne spécifie pas (et que l'on n'apprend pas) les corrélations entre les marginales.

intéressés par une *région de confiance* que par la simple collection de n intervalles de confiance. Une autre manière de voir CUCB est donc de considérer la région de confiance $\mathcal{C}_t$ qui lui est associée : il s'agit ni plus ni moins que du produit cartésien des intervalles de confiance qui étaient considérés par UCB. Voici une façon de décrire différemment CUCB en terme de région de confiance :

$$A_t \in \arg\max_{A \in \mathcal{A}} \max_{\boldsymbol{\mu} \in \mathcal{C}_t} \mathbf{e}_A^\intercal \boldsymbol{\mu}.$$

Nous avons donc deux aspects dans les méthodes optimistes : on peut soit regarder le *bonus d'exploration* qui est ajouté à l'estimation empirique (ici, le bonus est linéaire et vaut $\mathbf{e}_A^\intercal \left(\sqrt{\frac{2\log(t)}{N_{i,t-1}}} \right)_i$, on parle également de bonus de type ℓ_1), soit regarder la région de confiance autour du vecteur des moyennes empiriques $\overline{\boldsymbol{\mu}}_{t-1}$ (ici, la région est un hypercube, on parle également de région de type ℓ_∞).

Nous pouvons remarquer que le problème d'optimisation ci-dessus peut être résolu efficacement (c'est un problème d'optimisation linéaire sur $\mathcal{A}$) : CUCB est donc efficiente sur le plan statistique (parce qu'elle atteint la borne inférieure) *et* sur le plan computationnel.

Notons que la distribution $\mathbb{P}_{\mathbf{X}}$ choisie pour établir la borne inférieure ci-dessus est extrême : elle représente une situation où les bras appartenant à une même action sont parfaitement corrélés. Ce comportement extrême est également mis en évidence lorsque l'on regarde la région de confiance considérée par CUCB : on ne s'attend pas vraiment à avoir une zone de confiance ayant la forme d'un hypercube, mais plutôt ayant la forme d'une ellipsoïde, en prenant comme a priori une distribution Gaussienne multivariée. L'hypercube vient du fait que l'orientation et l'excentricité de l'ellipsoïde peuvent être arbitraire (voir Figure 1.2). Si, par exemple, nous nous limitons à la classe des distributions satisfaisant $\mathbb{P}_{\mathbf{X}} = \otimes_{i \in [n]} \mathbb{P}_{X_i}$ (ce qui signifie que les distributions des bras sont mutuellement indépendantes), alors la borne inférieure devient $\Omega(n \log(T)/\Delta)$ (car $\sigma_i^2 = m$ dans ce cas). Cela signifie que CUCB n'est plus efficiente sur le plan statistique, et que nous pouvons potentiellement gagner un facteur m dans la borne supérieure du regret. Combes et al. (2015) ont étudié ce problème, et ont donné une politique qui s'appuie effectivement sur l'indépendance des réalisations pour réduire la région de confiance utilisée. Plus précisément, ils

proposent une politique, appelée ESCB (pour *Efficient Sampling for Combinatorial Bandit*), qui jouent au tour t l'action

$$A_t \in \arg\max_{A\in\mathcal{A}} \mathbf{e}_A^\intercal \overline{\boldsymbol{\mu}}_{t-1} + \sqrt{\sum_{i\in A} \frac{f(t)}{N_{i,t-1}}},$$

où $f(t)$ est de l'ordre de $\mathcal{O}(\log(t))$. On remarque cette fois-ci qu'on a affaire à un bonus de type ℓ_2, correspondant à la région de confiance de type ℓ_2 (ici, l'ellipsoïde de confiance) suivante :

$$\mathcal{C}_t \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{ \boldsymbol{\xi} \in \mathbb{R}^n,\ \sum_{i\in[n]} N_{i,t-1}\xi_i^2 \leq f(t) \right\}.$$

Remarquons que l'ellipsoïde ci-dessus est *alignée par rapport aux axes* (ce qui est cohérent avec l'hypothèse d'indépendance). Degenne et Perchet (2016b) sont allés plus loin en considérant la classe des distributions $\mathbb{P}_{\mathbf{X}}$ qui sont $\mathbf{C}$-sous-Gaussiennes (multivariées), pour une matrice $\mathbf{C} \succeq 0$, i.e., telles que

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

Dans ce cas, ils ont construit une politique, OLS-UCB, nécessitant la connaissance d'une autre matrice $\boldsymbol{\Gamma} \succeq 0$ à coefficients positifs telle que $\boldsymbol{\Gamma} \succeq_+ \mathbf{C}$. Cette politique considère une ellipsoïde de confiance basée sur la matrice $\boldsymbol{\Gamma}$ (i.e., non alignée par rapport aux axes quand $\boldsymbol{\Gamma}$ n'est pas diagonale). Elle se réduit essentiellement à la politique ESCB dans le cas spécifique où la matrice $\boldsymbol{\Gamma}$ est diagonale. OLS-UCB a une borne sur le regret de l'ordre de

$$\mathcal{O}\left(\frac{\log T}{\Delta} \sum_{i\in[n]} \Gamma_{ii}\Big((1-\gamma)\log^2(m) + \gamma m\Big)\right),$$

où $\gamma \triangleq \max_{A\in\mathcal{A}} \max_{(i,j)\in A^2, i\neq j} \Gamma_{ij}/\sqrt{\Gamma_{ii}\Gamma_{jj}}$. On voit donc que ESCB, dans le cas de distributions mutuellement indépendantes, est efficiente sur le plan statistique (d'où son nom). En effet, on a alors $\gamma = 0$ dans la borne ci-dessus, donnant un regret pour ESCB de l'ordre de $\mathcal{O}(\log^2(m) n \log(T)/\Delta)$, qui est serré à un facteur polylogarithmique en m près. Degenne et Perchet (2016b) ont aussi montré que leur politique OLS-UCB était efficiente sur le plan statistique pour la classe des distributions sous-Gaussiennes multivariées considérées : ils ont en effet prouvé une borne inférieure de l'ordre de $\Omega\Big(\frac{n\log T}{\Delta}((1-\gamma)+\gamma m)\Big)$. On voit en particulier apparaître une interpolation entre deux régimes : pour γ proche de 0, on retrouve la même borne inférieure que dans le cas où les bras ont des distributions mutuellement indépendantes, et pour γ proche de 1, on retrouve la borne inférieure qui correspond au cas où les bras sont parfaitement corrélés.

Nous pouvons enfin remarquer que les politiques ESCB et OLS-UCB ne sont pas efficientes sur le plan computationnel en général (car le problème combinatoire à résoudre à chaque tour n'est plus linéaire, et est même NP-difficile en général (Atamtürk and Gómez, 2017)).

1.2.2 Un bref aperçu sur les récompenses non linéaires et les contraintes de budget

Généralement, on suppose que la récompense espérée lorsque l'agent choisit une action A est de la forme $r(A; \boldsymbol{\mu}^*)$. Si telle est le cas, la fonction r est appellée *fonction de récompense*. Elle est dite linéaire lorsque $r(A; \boldsymbol{\mu}^*) = \mathbf{e}_A^\intercal \boldsymbol{\mu}^*$. On suppose ici, pour simplifier, que pour un vecteur $\boldsymbol{\mu}$ fixé, la fonction $A \mapsto r(A; \boldsymbol{\mu})$ peut être maximisé exactement et efficacement sur $\mathcal{A}$. Nous avons déjà mentionné que cette thèse ne se concentrait pas sur les fonctions de récompense linéaires. Néanmoins, la plupart des fonctions de récompense que nous rencontrerons ont un comportement semblable au cas des récompenses linéaires. Pour être plus précis, nous verrons qu'il existe fréquemment un contrôle en norme ℓ_1 sur la déviation de la récompense lorsque le paramètre que l'on désire apprendre (ici la moyenne) varie. Par exemple, une hypothèse usuelle, considérée par Wang et Chen (2017), est

$$\left| r(A; \boldsymbol{\mu}) - r(A; \boldsymbol{\mu}') \right| \leq \left\| \mathbf{e}_A \odot (\boldsymbol{\mu} - \boldsymbol{\mu}') \right\|_1. \tag{1.2}$$

Lorsque cette hypothèse est vérifiée, les approches optimistes ci-dessus restent valables (c'est-à-dire que les bornes sur le regret sont du même ordre) en considérant, dans les politiques, le problème d'optimisation

$$\max_{A \in \mathcal{A},\ \boldsymbol{\mu} \in C_t} r(A; \boldsymbol{\mu})$$

afin de choisir A_t. Une difficulté des récompenses non linéaires dans les approches optimistes est que la quantité $\max_{\boldsymbol{\mu} \in C_t} r(A; \boldsymbol{\mu})$ peut ne pas être simple à calculer (et encore moins à optimiser sur $A \in \mathcal{A}$). Néanmoins, il existe des cas où c'est possible. Par exemple, si nous avons une propriété de *croissance* (composante par composante) de la fonction $\boldsymbol{\mu} \mapsto r(A; \boldsymbol{\mu})$, alors, pour une région de confiance de type ℓ_∞, on peut directement appliquer le vecteur des estimations optimistes au lieu de la vraie moyenne afin de sélectionner l'action A_t à prendre (Chen, Wang, et Yuan, 2013, 2016), il en résulte donc que cette politique est efficiente sur le plan computationnel.

Un cas particulièrement intéressant de fonction de récompense non linéaire apparaît lorsque nous considérons un problème de *bandit combinatoire à budget* (Xia, Qin, et al., 2016). Pour expliquer brièvement ce cadre, il n'y a plus d'horizon T, mais plutôt un budget B, et chaque action qui est prise lors d'un tour est coûteuse. Nous verrons que ce cadre peut être essentiellement résolu de la même manière que le problème standard. Une différence notable est que la fonction objectif à optimiser au sein de chaque tour n'est plus une fonction de récompense, mais un rapport entre une fonction de récompense et une fonction de coût. Ce type de problème est donc à l'origine de nombreuses fonctions objectifs non linéaires dans les problèmes CMAB, même si les fonctions de récompense et de coût étaient initialement linéaires.

1.2.3 Bras à déclenchement probabiliste

Dans le cadre typique du problème CMAB, l'ensemble des bras déclenchés et l'action joué par l'agent sont confondus. Plus précisément, l'agent sélectionne une action à jouer à chaque tour, déclenchant un ensemble de bras, et les réalisations de ces bras sont alors observées. Une généralisation intéressante, appelé *bras à déclenchement probabiliste* (que nous abrégeons en CMAB-T dans cette thèse), a été introduite par Chen, Wang, et Yuan (2016) et Wang et Chen (2017). L'idée est que l'action sélectionnée par l'agent n'est plus réduite à un ensemble de bras observés, mais appartient

à un espace extérieur. Après qu'une action soit sélectionnée, certains bras sont déclenchés de manière probabiliste. Ce cadre a notamment été utilisé afin d'aborder des problèmes comme les *bandits en cascade* (portant sur la sélection des pages web à afficher pour une requête sur un moteur de recherche, Kveton, Szepesvari, et al. (2015) et Kveton, Wen, Ashkan, and Szepesvari (2015a)) ou encore la *maximisation d'influence incrémentielle* (portant sur la sélection d'influenceurs dans un réseau social, Chen, Wang, et Yuan (2016), Wen, Kveton, Valko, et al. (2017) et Wang et Chen (2017)).

Une hypothèse très pratique pour conserver la borne sur le regret obtenue plus haut pour les régions de type ℓ_∞ est de considérer la même relation de régularité que (1.2), mais en pondérant chaque terme (correspondant à un bras) dans la norme par la probabilité d'observer le bras correspondant en choisissant l'action. Cette hypothèse est vérifiée notamment dans les problèmes de bandit en cascade et de maximisation d'influence incrémentielle (Wang et Chen (2017)).

1.2.4 Problématiques de la thèse et contributions

Après l'état des lieux qui précède, de nombreuses questions se posent, et chercher à y répondre constitue les objectifs de cette thèse. De manière générale, on s'intéressera à améliorer l'efficience des politiques, tant du point de vue computationnel que du point de vue statistique. Plus précisément, nous porterons un intérêt particulier pour les questions suivantes :

- Pouvons-nous espérer implémenter une version efficiente des politiques ESCB et OLS-UCB ? Existe-t-il des alternatives qui sont efficientes tant sur le plan computationnel que sur le plan statistique (dans le cas de distributions indépendantes, ou sous-Gaussiennes multivariées) ?

- Est-il possible d'associer l'analyse ℓ_2 avec le cadre des contraintes de budget ou des bras à déclenchement probabiliste ? Quel genre de borne sur le regret pouvons-nous obtenir ? Une telle analyse peut-elle contribuer à améliorer les politiques existantes (et le regret qu'elles suscitent) pour des problèmes connus, comme la maximisation d'influence incrémentielle ?

- Pouvons-nous optimiser l'analyse de Degenne et Perchet (2016b) ? Par exemple, pouvons-nous assouplir l'hypothèse selon laquelle la matrice de sous-Gaussianité $\mathbf{\Gamma}$ est connue ? Existe-t-il une alternative à la famille des variables aléatoires sous-Gaussiennes multivariées ? Est-il possible d'expliciter le comportement d'interpolation dans le regret, par exemple en remplaçant $\Gamma_{ii}(1-\gamma+\gamma m)$ par la quantité plus faible $\max_{A\in\mathcal{A}:\, i\in A} \sum_{j\in A} \Gamma_{ij}$?

Nous allons maintenant présenter le contenu de la thèse, chapitre par chapitre, en mettant l'accent sur les aspects liés aux questions soulevées ci-dessus.

Chapitre 2, Bandit Manchot Stochastique Nous passons d'abord en revue les résultats de base du problème classique des bandits à plusieurs bras. Certains d'entre eux seront utiles plus tard dans le contexte combinatoire.

Chapitre 3, Cadre Général pour les Observations Semi-Bandits Nous formalisons ici le cadre des semi-bandits stochastiques combinatoires avec bras à déclenchement probabiliste (qui, on le rappelle, englobe le cadre habituel). Nous donnons ensuite une multitude d'applications de ce cadre que l'on trouve dans la littérature. Enfin, nous fournissons des résultats techniques généraux qui seront utiles tout

au long de la thèse pour prouver des bornes supérieures sur le regret des politiques. Plus précisément, les théorèmes énoncés dépendent du type d'erreur que l'on veut contrôler (qui n'est rien d'autre que le type de bonus que l'on utilise). Pour l'analyse ℓ_∞ (i.e., basée sur une région de confiance de type ℓ_∞, ou d'une manière équivalente sur un bonus de type ℓ_1), nous avons un théorème qui améliore légèrement (et surtout simplifie grandement) l'analyse ℓ_∞ de Wang et Chen (2017). Pour ce qui est de l'analyse ℓ_2, nous fournissons plusieurs nouveaux résultats, étendant le travail de Degenne et Perchet (2016b), et surpassant l'analyse ℓ_∞ (gagnant un facteur m, à un facteur polylogarithmique près). Ce faisant, nous parvenons à associer l'analyse ℓ_2 avec le cadre des bras à déclenchement probabiliste. Ce chapitre contient quelques résultats non publiés, et quelques améliorations de résultats publiés dans :

- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020a) "Budgeted Online Influence Maximization", dans 37th International Conference on Machine Learning (ICML).
- Pierre Perrault, Vianney Perchet, Michal Valko (2020) "Covariance-adapting algorithm for semi-bandits with application to sparse rewards", dans 33rd Conference on Learning Theory (COLT).
- Pierre Perrault, Etienne Boursier, Vianney Perchet, Michal Valko (2020) "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits", dans 34th Conference on Neural Information Processing Systems (NeurIPS).

Chapitre 4, Un Exemple de Problème de type CMAB-T : Recherche-et-Arrêt Séquentiels Nous présentons ici un nouvel exemple de problème qui se situe à la fois dans le cadre du bras à déclenchement probabiliste et dans le cadre à budget. Nous appelons ce problème *recherche-et-arrêt séquentiels*. Il est décrit de la manière suivante : on considère un graphe acyclique dirigé, où chaque nœud est un bras associé à un coût. Il y a un objet qui est caché au hasard dans l'un des nœuds (selon une distribution inconnue à apprendre). Dans un tour, l'agent peut inspecter des nœuds, choisis un par un, avec la contrainte qu'un nœud est accessible si ses voisins entrants ont déjà été inspectés. Il peut décider à tout moment d'arrêter sa recherche, et ainsi passer à une nouvelle instance indépendante (où l'objet est replacé au hasard). L'agent dispose d'un budget initial, et chaque fois qu'un nœud est inspecté, son prix est payé. L'objectif de l'agent est de trouver un nombre maximal d'objet (en espérance). Nous fournissons une politique basée sur une approche ℓ_∞, en décrivant notamment un stratégie quasi-optimale et efficace pour le problème "hors-ligne" (i.e., lorsque la distribution est connue par l'agent, qui doit alors se focaliser sur l'exploitation). Ce chapitre est adapté de la publication suivante :

- Pierre Perrault, Vianney Perchet, Michal Valko (2019b) "Finding the bandit in a graph: Sequential search-and-stop", dans 22nd International Conference on Artificial Intelligence and Statistics (AIStats).

Chapitre 5, La Structure de l'Incertitude Nous commençons par rappeler les deux principales méthodes pour obtenir des zones de confiance ℓ_2 : *la méthode de Laplace*, et *l'argument de la couverture*. Nous étudions ensuite plus en profondeur la structure des bonus ℓ_2, et nous prouvons qu'en tant que fonctions sur $\mathcal{P}([n])$, elles sont sous-modulaires. Nous nous appuyons ensuite sur cette structure pour proposer plusieurs algorithmes d'approximation afin de maximiser la somme d'une fonction linéaire ($A \mapsto \mathbf{e}_A^\top \overline{\boldsymbol{\mu}}_{t-1}$) et d'une fonction sous-modulaire (le bonus d'exploration). Ceci

nous permet d'obtenir des approximations de ESCB pour des contraintes de *matroïde* (un certain type d'espace d'action $\mathcal{A}$ généralisant la contrainte sur le cardinal). Ces approximations sont efficientes sur le plan computationnel et ne dégradent pas l'ordre du regret. Nous fournissons également une extension de cette méthode au cadre à budget. Ce chapitre est basé sur la publication suivante :

- Pierre Perrault, Vianney Perchet, Michal Valko (2019a) "Exploiting structure of uncertainty for efficient matroid semi-bandits", dans 36th International Conference on Machine Learning (ICML).

Chapitre 6, Maximisation d'Influence Incrémentielle Budgétisée Ce chapitre vise à approfondir l'étude d'un autre exemple bien connu de problème impliquant des bras à déclenchement probabiliste, à savoir la maximisation d'influence incrémentielle. Pour rappel, dans ce problème, un agent apprend activement à connaître un réseau social en interagissant avec lui de manière répétée, en essayant de trouver les meilleurs ensembles d'influenceurs. Nous incorporons un cadre à budget à ce problème, en considérant le coût total d'une campagne publicitaire au lieu de la contrainte (par tour) de cardinalité. Notre approche modélise ainsi mieux le contexte du monde réel où le coût des influenceurs varie et où les publicitaires veulent trouver le meilleur rapport qualité-prix pour leur budget global. Nous proposons des politiques de types ℓ_∞ et ℓ_2, en utilisant la sous-modularité d'une borne supérieure sur les bonus ℓ_2. Ce chapitre est adapté de la publication et de la prépublication suivantes :

- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020a) "Budgeted Online Influence Maximization", dans 37th International Conference on Machine Learning (ICML).
- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020b) "On the Approximation Relationship between Optimizing Ratio of Sub-modular (RS) and Difference of Submodular (DS) Functions", en cours de révision.

Chapitre 7, Politique Adaptative à la Covariance Le but ici est d'offrir une solution plus pratique, et plus précise pour construire une région de confiance ℓ_2. En effet, celles que nous avons vues jusqu'à présent sont soit basées sur l'indépendance mutuelle des réalisations, soit dues à la connaissance irréaliste d'une matrice de sous-Gaussianité $\mathbf{\Gamma}$. Nous considérons une nouvelle famille générale de distributions *sous-exponentielles* paramétrée par la matrice de covariance, supposée inconnue. Cette famille contient les distributions à support borné et les Gaussiennes. Nous prouvons une nouvelle borne inférieure sur le regret de cette famille, qui est paramétrée par la matrice de covariance, et non plus la matrice de sous-Gaussianité. Nous construisons ensuite un algorithme qui utilise des estimations de la covariance, et fournissons une analyse asymptotique du regret, atteignant la borne inférieure. Enfin, nous appliquons et étendons nos résultats à la famille des réalisations *parcimonieuses*, qui a des applications dans de nombreux systèmes de recommandation. Ce chapitre est adapté de la publication suivante :

- Pierre Perrault, Vianney Perchet, Michal Valko (2020) "Covariance-adapting algorithm for semi-bandits with application to sparse rewards", dans 33rd Conference on Learning Theory (COLT).

Chapitre 8, Efficience Statistique et Computationnelle de l'Echantillonnage de Thompson Dans ce chapitre, nous étudions une alternative intéressante aux

méthodes optimistes envisagées jusqu'à présent, à savoir *l'échantillonnage de Thompson*. Outre sa supériorité empirique (Chapelle et Li, 2011), l'échantillonnage de Thompson est intéressant dans le contexte des bandits combinatoires car l'action à jouer est facile à calculer : il ne faut considérer ni bonus, ni maximum sur une région de confiance, car le paramètre à utiliser dans la fonction objectif est simplement pris au hasard, selon un a priori bien choisi. L'échantillonnage de Thompson pourrait donc répondre à la question de l'existence d'une politique efficace (computationnellement) avec un regret asymptotique optimal (à un facteur polylogarithmique en m près), qui est encore ouverte pour de nombreuses familles de distributions, incluant les réalisations mutuellement indépendantes, et plus généralement la famille des réalisations sous-Gaussiennes multivariées. Nous proposons de répondre à la question ci-dessus pour ces deux familles en analysant des variantes de la politique *Combinatorial Thompson Sampling* (CTS). Pour des réalisations mutuellement indépendantes dans $[0, 1]$, nous proposons une borne serrée pour le regret de CTS, en utilisant la loi bêta comme a priori. Nous examinons ensuite le cadre plus général des réalisations sous-Gaussiennes multivariées et proposons une borne serrée pour le regret de CTS à l'aide d'un a priori Gaussien. Ce dernier résultat nous donne une alternative efficiente à la politique ESCB. Ce chapitre est adapté de la publication suivante :

- Pierre Perrault, Etienne Boursier, Vianney Perchet, Michal Valko (2020) "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits", dans 34th Conference on Neural Information Processing Systems (NeurIPS).

Chapitre 9, Conclusions et Perspectives Ce chapitre tire les conclusions de la thèse et fournit quelques orientations pour les futurs travaux.

1.3 Presentation of the thesis content (in English)

This thesis studies sequential combinatorial optimization problems in a stochastic environment, with observations that we call "semi-bandit". The terminology "semi-bandit" refers to the famous stochastic multi-armed bandit problem. Bandit problems are an important class of sequential optimization problems and have been widely studied within the field of statistics and machine learning. They were originally introduced in the seminal paper of Robbins (1952) to study sequential experimental designs. The classical version of the problem is formulated as a system of n arms (or machines). An *agent* must repeatedly select an arm from the n. Each arm i has an unknown probability distribution $\mathbb{P}_{X_i}$, of unknown mean μ_i^*, where the variable X_i encodes a *reward*. After selecting an arm, the agent observes an independent realization (also called outcome) of the corresponding reward distribution. Each decision is based on past decisions and observed rewards. The agent's task is to sequentially play the arms in order to maximize the expectation of the cumulative reward. The agent must play by balancing exploitation and exploration. Indeed, the arms with the highest observed rewards must be selected often, while all arms must be explored to know their mean rewards. Equivalently, the performance of the agent (or policy) can be evaluated by its regret R_T, defined as the expectation of the gap on T rounds ($T \in \mathbb{N}^*$ is called the horizon) between the reward accumulated by the policy and that accumulated by a policy that always selects the best arm. Regret can be rewritten as

$$R_T \triangleq \sum_{i \in [n]} \mathbb{E}[N_{i,T}] \Delta_i,$$

where $\Delta_i \triangleq \max_j \mu_j^* - \mu_i^*$ is the difference in expectation between the optimal arm and the arm i, and where $N_{i,t}$ is the number of times the arm i has been pulled until the round $t \in \mathbb{N}^*$. The notion of regret thus quantifies the loss due to the need to learn the mean rewards of the different arms.

The problem defined above is an example of a reinforcement learning problem. Indeed, a multi-armed bandit problem can be seen as a Markov decision process with a single state. Let's mention in passing that the name of the problem comes from imagining a player facing a row of slot machines, having to decide which machines to play, how many times to play each machine, and in what order to play them. Of course, this is only a model, and in practice this problem addresses any kind of situation where an agent simultaneously tries to acquire new knowledge (exploration) and to optimize its decisions on the basis of existing knowledge (exploitation). There are many practical applications of the bandit model, for example:

- Clinical trials that study the effects of different experimental treatments while minimizing the loss of human life (Thompson, 1933; Gittins, Weber, and Glazebrook, 1989; Berry and Fristedt, 1985).
- The design of financial portfolios (Hoffman, Brochu, and Freitas, 2011).
- Decision-making for dynamic spectrum access in the context of cognitive radios (Lai, Jiang, and Poor, 2008).

Multi-armed bandit has been studied by Lai and Robbins (1985). They proved an asymptotic lower bound on the regret R_T. This bound indicates that any "reasonable" policy (in a certain sense) must have at least a logarithmic regret with respect to the horizon T. It thus provides a fundamental performance limit that no "reasonable" policy can beat. The constant in the lower bound is linear in n, the number of arms. More precisely, it is worth the sum on the sub-optimal arms i of the difference Δ_i times the inverse of the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the distribution of the arm i and that of the optimal arm. Roughly, $\log(T)$ times the inverse of the KL represents the number of rounds needed to distinguish the arm i from the optimal one, giving a clear interpretation of the lower bound. In the case of Gaussian distributions, the inverse of the KL is lower bounded (up to a multiplicative constant) by the variance σ_i^2 of the arm i divided by Δ_i^2. To simplify, the lower bound thus becomes

$$\Omega\left(\log(T) \sum_{i \in [n],\ \Delta_i > 0} \frac{\sigma_i^2}{\Delta_i}\right). \tag{1.3}$$

It is not surprising that the variance appears, as it can be seen as a measure of the uncertainty we have in our samples: the higher the variance, the more difficult the estimation, and therefore the higher the regret. Lai and Robbins (1985) also developed policies for some reward distributions and showed that they asymptotically reach the lower bound. Indeed, these policies have an upper bound on their regret which is logarithmic in T, with a constant of the same order as in the lower bound. In the case where the arms have distributions σ_i^2-sub-Gaussian, the strategy UCB (Upper Confidence Bound (Auer, Cesa-Bianchi, and Fischer, 2002)) asymptotically reaches the bound (1.3). It consists in choosing, in round t, the arm i_t maximizing the upper confidence bound $\overline{\mu}_{i,t-1} + \sqrt{2\log(t)/N_{i,t-1}}$, where $\overline{\mu}_{i,t-1}$ is the empirical average of the rewards of the arm i. This is a strategy based on the *optimistic in face of uncertainty* principle, because it chooses the arm that would be best if the means were worth their optimistic estimates.

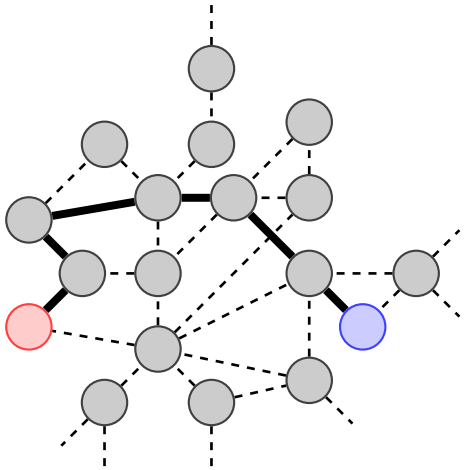

FIGURE 1.3: Example of a network (graph) where each edge is an arm. The node in blue represents the origin, and the node in red represents the destination. The edges represented by bold lines form a path between the origin and the destination: it is therefore an action that the agent can decide to play during a round. Doing so, each arm of the path produces one outcome, and the reward can then be the sum of the outcomes of the path.

Stochastic combinatorial bandit To come back to the subject of this thesis, i.e., sequential combinatorial optimization problems, we can extend the above-mentioned framework: the agent can now, during a round, choose several arms instead of one (with possibly some constraints on the possible sets of arms). The reward obtained depends on the individual outcome of each chosen arm. This is well adapted to applications where the actions available to the agent belong to a combinatorial space and are built from several small actions (the arms). Let us mention for example the problem of the shortest path between two points of a network (which has been studied in particular by Liu and Zhao (2012) and Talebi, Zou, et al. (2013)). Each edge of the network is an arm associated to a probability distribution, modelling for example the traffic flow on this edge. The agent chooses at each round a path connecting the two points, and is rewarded with an independent realization of the overall traffic flow of the chosen path. This overall traffic flow is defined according to the context, and can be for example the sum of the path's edge flows (see Figure 1.3). This particular case where the reward is the sum of the outcomes of the chosen arms corresponds to a sequential combinatorial optimization problem with a *linear* objective function, and it is the most common scenario among these types of problems. It is studied, for example, by Abernethy, Hazan, and Rakhlin (2008), Dani, Hayes, and Kakade (2008), and Bubeck, Cesa-Bianchi, and Kakade (2012). In general, the combinatorial extension of the bandit problem is known in the literature as *stochastic combinatorial bandit*. It implies (at least) two distinct frameworks concerning the observations available to the agent on the outcomes of the arms at each round:

- *Semi-bandit feedback*: all the outcomes of the selected arms are observed by the agent.
- *Bandit feedback*: only the reward (which is a function of the outcomes of the chosen arms) is observed.

As we have already mentioned, we are interested here in the first framework, commonly called *stochastic combinatorial semi-bandits* (which we will abbreviate to CMAB in this thesis, for *combinatorial multi-armed bandits*, the "semi" being omitted for sake of brevity). Since the agent has more observations, this assumption may be more difficult to satisfy in practice, but may lead to better policies. Note also that the second framework is in fact a special case of the bandit problem introduced above, where the number of arms is combinatorial, but where a certain structure is specified between rewards.

1.3.1 Semi-bandit feedback

Here we formalize the CMAB problem. The outcome of the arm $i \in [n]$ in round t is denoted as $X_{i,t} \in \mathbb{R}$. The random vectors $(\mathbf{X}_t)_{t\geq 1}$ of $\mathbb{R}^n$ are i.i.d., with an unknown distribution $\mathbb{P}_{\mathbf{X}}$, of mean $\boldsymbol{\mu}^* \in \mathbb{R}^n$. Nevertheless, unless otherwise stated, X_i and X_j for two distinct arms $i \neq j$ can be arbitrarily correlated. The action (also called *super-arm*) selected by the agent at round t (i.e., the set of arms selected) is denoted A_t. The set of possible super-arms is called *action space*, and is noted $\mathcal{A}$. It is a fixed subset of $\mathcal{P}([n])$, such that each of its elements A is a subset of at most m arms. After selecting an arm A_t in round t, the agent receives as feedback $\mathbf{e}_{A_t} \odot \mathbf{X}_t$. Although we consider a more general framework in the thesis, in order to simplify the explanations, we limit ourselves here to the case of a linear reward function: the reward at round t is therefore $\mathbf{e}_{A_t}^\intercal \mathbf{X}_t$. We also assume that $\mathcal{A}$ is such that the optimization of linear functions can be done efficiently (in time polynomial in n).

The objective is to identify a policy that maximizes the expected cumulative reward over the T rounds. The expectation is taken on the randomness of the rewards and on the possible randomness of the policy followed by the agent (an action can be chosen at random). In an equivalent way, we aim at designing a policy π that minimizes the regret, defined as:

$$R_T(\pi) \triangleq \max_{A\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} (\mathbf{e}_A - \mathbf{e}_{A_t})^\intercal \mathbf{X}_t\right].$$

Contrary to the classic bandit problem, there is a notion of similarity between actions. Indeed, two "close" actions, i.e., having many arms in common, tend to have a "close" reward. This structure can be exploited in the learning process, in order to counterbalance the combinatorial size of the action space. This results in the design of efficient super-arm selection policies, based on the principle of optimism in face of the uncertainty, whose regret is of the order of $\mathcal{O}(n \cdot m \log(T)/\Delta)$ (Kveton, Wen, Ashkan, and Szepesvari, 2015b), where $\Delta > 0$ is the minimum difference between the mean of an optimal super-arm and the mean of a sub-optimal super-arm. These policies asymptotically reach a lower bound: considering the action space $\mathcal{A} = \{A_1, \ldots, A_{n/m}\}$ with $A_k = \{(k-1)m+1, \ldots, km\}$, and posing for all k, $X_{(k-1)m+1} = \cdots = X_{km}$, drawn according to a Gaussian of variance 1, we reduce the problem to a classical bandit problem, allowing us to apply the bound (1.3), with $\sigma_i^2 = m^2$. The best-known policy of this kind is CUCB (*Combinatorial Upper Confidence Bound*), which is heavily influenced by UCB. Indeed, it plays at round t the action

$$A_t \in \arg\max_{A\in\mathcal{A}} \mathbf{e}_A^\intercal \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{2\log(t)}{N_{i,t-1}}}\right)_i,$$

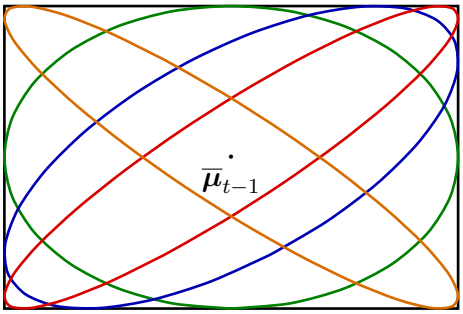

FIGURE 1.4: The hypercube type confidence region (in black) corresponds to a Gaussian prior for each marginal. However, the correlation between these marginals can be arbitrary: here are several examples, from very negatively correlated (in orange) to very positively correlated (in red), through independent (in green). We can see that the whole hypercube is likely to contain the true mean $\boldsymbol{\mu}^*$ as long as we do not specify (and learn) the correlations between the marginals.

in other words, the optimistic estimates are the same as for UCB. Note that in CMAB problems, the entire joint distribution of the random vector $\mathbf{X}$ matters, unlike in classical bandits where only the marginals are sufficient to characterize a problem instance. Thus, when estimating the vector of means $\boldsymbol{\mu}^*$, we are more interested in a *confidence region* than just a collection of n confidence intervals. Another way to look at CUCB is therefore to consider the confidence region $\mathcal{C}_t$ associated with it: it is no more than the Cartesian product of the confidence intervals that were considered by UCB. Here is a way of rewriting CUCB in terms of confidence region:

$$A_t \in \arg\max_{A\in\mathcal{A}} \max_{\boldsymbol{\mu}\in\mathcal{C}_t} \mathbf{e}_A^\intercal \boldsymbol{\mu}.$$

We thus have two view points in the optimistic methods: We can either look at the exploration bonus that is added to the empirical estimate (here, the bonus is linear and is worth $\mathbf{e}_A^\intercal \left(\sqrt{\frac{2\log(t)}{N_{i,t-1}}}\right)_i$, we also say the bonus is of type ℓ_1), or look at the confidence region around the vector of empirical means $\overline{\boldsymbol{\mu}}_{t-1}$ (here, the region is a hypercube, we also say the region is of type ℓ_∞).

We can notice that the above optimization problem can be solved efficiently (it is a linear optimization problem on $\mathcal{A}$): CUCB is therefore statistically efficient (because it reaches the lower bound) *and* computationally efficient.

Note that the distribution $\mathbb{P}_{\mathbf{X}}$ chosen to establish the above lower bound is extreme: it represents a situation where the arms belonging to the same action are perfectly correlated. This extreme behavior is also highlighted when we look at the confidence region considered by CUCB: we do not really expect to have a confidence zone having the shape of a hypercube, but rather having the shape of an ellipsoid, taking as prior a multivariate Gaussian distribution. The hypercube comes from the fact that the orientation and eccentricity of the ellipsoid can be arbitrary (see Figure 1.4). If, for example, we limit ourselves to the class of distributions that satisfy $\mathbb{P}_{\mathbf{X}} = \otimes_{i\in[n]}\mathbb{P}_{X_i}$ (which means that the distributions of the arms are mutually independent), then the lower bound becomes $\Omega(n\log(T)/\Delta)$ (because $\sigma_i^2 = m$ in this case). This means that CUCB is no longer statistically efficient, and that we can potentially gain a factor m in the regret upper bound. Combes et al. (2015) have studied this problem, and have come up with a policy that indeed relies on the independence of the outcomes to reduce the confidence region used. Specifically, they proposed a

policy, called ESCB (for *Efficient Sampling for Combinatorial Bandit*), which plays at round t the action

$$A_t \in \arg\max_{A\in\mathcal{A}} \mathbf{e}_A^\mathsf{T}\overline{\boldsymbol{\mu}}_{t-1} + \sqrt{\sum_{i\in A}\frac{f(t)}{N_{i,t-1}}},$$

where $f(t)$ is of order $\mathcal{O}(\log(t))$. This time we notice that we're dealing with an ℓ_2 bonus, corresponding to the following ℓ_2 confidence region (here, a confidence ellipsoid):

$$\mathcal{C}_t \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{\boldsymbol{\xi}\in\mathbb{R}^n,\ \sum_{i\in[n]} N_{i,t-1}\xi_i^2 \le f(t)\right\}.$$

Note that the above ellipsoid is *aligned with respect to the axes* (which is consistent with the independence assumption). Degenne and Perchet (2016b) went further by considering the class of distributions $\mathbb{P}_\mathbf{X}$ which are $\mathbf{C}$-sub-Gaussian (multivariate), for a matrix $\mathbf{C}\succeq 0$, i.e., such that

$$\forall\boldsymbol{\lambda}\in\mathbb{R}^n,\ \mathbb{E}\left[e^{\boldsymbol{\lambda}^\mathsf{T}(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \le e^{\boldsymbol{\lambda}^\mathsf{T}\mathbf{C}\boldsymbol{\lambda}/2}.$$

In this case, they provided a policy, OLS-UCB, requiring knowledge of another matrix $\boldsymbol{\Gamma}\succeq 0$ with non-negative coefficients such that $\boldsymbol{\Gamma}\succeq_+\mathbf{C}$. This policy considers a confidence ellipsoid based on the matrix $\boldsymbol{\Gamma}$ (i.e., not aligned with respect to the axes when $\boldsymbol{\Gamma}$ is not diagonal). It essentially boils down to the policy ESCB in the specific case where the matrix $\boldsymbol{\Gamma}$ is diagonal. OLS-UCB has a regret bound of order

$$\mathcal{O}\left(\frac{\log T}{\Delta}\sum_{i\in[n]}\Gamma_{ii}\Big((1-\gamma)\log^2(m)+\gamma m\Big)\right),$$

where $\gamma \triangleq \max_{A\in\mathcal{A}}\max_{(i,j)\in A^2, i\neq j}\Gamma_{ij}/\sqrt{\Gamma_{ii}\Gamma_{jj}}$. So we can see that ESCB, in the case of mutually independent distributions, is statistically efficient (hence its name). Indeed, we then have $\gamma=0$ in the above bound, giving a regret for ESCB of the order of $\mathcal{O}(\log^2(m)n\log(T)/\Delta)$, which is tight up to a polylogarithmic factor in m. Degenne and Perchet (2016b) also showed that their policy was statistically efficient for the class of multivariate sub-Gaussian distributions under consideration. They proved a lower bound on the order of $\Omega\Big(\frac{n\log T}{\Delta}\big((1-\gamma)+\gamma m\big)\Big)$. In particular, an interpolation between two regimes appears: for γ close to 0, we find the same lower bound as in the case where the arms have mutually independent distributions, and for γ close to 1, we find the lower bound corresponding to the case where the arms are perfectly correlated.

We can finally notice that the policies ESCB and OLS-UCB are not computationally efficient in general (because the combinatorial problem to be solved at each round is no longer linear, and is even NP-hard in general (Atamtürk and Gómez, 2017)).

1.3.2 A brief look at non-linear rewards and budget constraints

Generally, it is assumed that the expected reward when the agent chooses an action A is in the form $r(A;\boldsymbol{\mu}^*)$. If this is the case, the function r is called *reward function*. It is linear when $r(A;\boldsymbol{\mu}^*)=\mathbf{e}_A^\mathsf{T}\boldsymbol{\mu}^*$. It is assumed here, for simplicity, that for a fixed vector $\boldsymbol{\mu}$, the function $A\mapsto r(A;\boldsymbol{\mu})$ can be maximized exactly and efficiently on $\mathcal{A}$. We have already mentioned that this thesis did not focus solely on linear reward functions. Nevertheless, most of the reward functions we will encounter behave similarly to

linear rewards. To be more precise, we will see that there is frequently an ℓ_1-norm control on the deviation of the reward when the parameter we want to learn (here the mean) varies. For example, a common assumption considered by Wang and Chen (2017) is

$$|r(A;\boldsymbol{\mu}) - r(A;\boldsymbol{\mu}')| \leq \|\mathbf{e}_A \odot (\boldsymbol{\mu} - \boldsymbol{\mu}')\|_1. \tag{1.4}$$

Under this assumption, the optimistic approaches above remain valid (i.e. the bounds on the regret are of the same order) by considering, in policies, the optimization problem

$$\max_{A\in\mathcal{A},\ \boldsymbol{\mu}\in C_t} r(A;\boldsymbol{\mu})$$

in order to select A_t. One difficulty with non-linear rewards in optimistic approaches is that the quantity $\max_{\boldsymbol{\mu}\in\mathcal{C}_t} r(A;\boldsymbol{\mu})$ may not be simple to compute (and even less to optimize over $A \in \mathcal{A}$). Nevertheless, there are cases where this is possible. For example, if we have a *monotonicity* property (component-wise) of the function $\boldsymbol{\mu} \mapsto r(A;\boldsymbol{\mu})$, then, for a confidence region of type ℓ_∞, we can directly plug the vector of optimistic estimates instead of the true mean in order to select the action A_t (Chen, Wang, and Yuan, 2013; Chen, Wang, and Yuan, 2016), the result is that this policy is computationally efficient.

A particularly interesting case of a non-linear reward function arises when we consider a *budgeted combinatorial bandit* problem (Xia, Qin, et al., 2016). To briefly explain this framework, there is no longer a time horizon T, but rather a budget B, and every action that is taken in a round is costly. We will see that this framework can be essentially solved in the same way as the standard problem. A notable difference is that the objective function to be optimized within each round is no longer a reward function, but a ratio between a reward function and a cost function. This type of problem is therefore at the origin of many non-linear objective functions in CMAB problems, even if the reward and cost functions were initially linear.

1.3.3 Probabilistically triggered arms

In the typical CMAB problem, the set of triggered arms and the action played by the agent coincide. Specifically, at each round, the agent selects an action to play, which triggers a set of arms, and the outcomes of these arms are then observed. An interesting generalization, called *probabilistically triggered arms* (which we abbreviate to CMAB-T in this thesis), was introduced by Chen, Wang, and Yuan (2016) and Wang and Chen (2017). The idea is that the action selected by the agent is no longer reduced to a set of observed arms, but belongs to an external space. After an action is selected, some arms are triggered probabilistically. In particular, this framework has been used to address problems such as *Cascading bandits* (about the selection of web pages to be displayed for a search engine query (Kveton, Szepesvari, et al., 2015; Kveton, Wen, Ashkan, and Szepesvari, 2015a)) or online influence maximization (about selecting influencers in a social network (Chen, Wang, and Yuan, 2016; Wen, Kveton, Valko, et al., 2017; Wang and Chen, 2017)). A very practical assumption so that the above regret bound still (for ℓ_∞ type regions) holds is to consider the same smoothness relation as (1.4), but by weighting each term (corresponding to an arm) in the norm by the probability of observing the corresponding arm when choosing the action. This assumption is verified in particular in the problems of cascading bandit and online influence maximization (Wang and Chen, 2017).

1.3.4 Challenges of the thesis and contributions

After the above overview, several questions can be asked, and trying to answer them constitutes the objectives of this thesis. In general, the focus will be on improving the efficiency of policies, both from a computational and statistical point of view. Specifically, we will be particularly interested in the following questions:

- Can we hope to implement an efficient version of the ESCB and OLS-UCB policies? Are there alternatives that are computationally as well as statistically efficient (in the case of independent, or multivariate sub-Gaussian distributions)?
- Is it possible to associate the ℓ_2 analysis with the budgeted or the probabilistically triggered arms frameworks? What kind of bound on the regret can we get? Can such an analysis help improve existing policies (and the regret they have) for known problems, such as online influence maximization?
- Can we optimize the analysis from Degenne and Perchet (2016b)? For example, can we relax the assumption that the sub-Gaussianity matrix $\mathbf{\Gamma}$ is known? Is there an alternative to the family of multivariate sub-Gaussian random variables? Is it possible to explicit the interpolation behavior in the regret, for example by replacing $\Gamma_{ii}(1-\gamma+\gamma m)$ by the smaller quantity $\max_{A\in\mathcal{A}:\, i\in A} \sum_{j\in A} \Gamma_{ij}$?

We will now present the contents of the thesis, chapter by chapter, focusing on aspects related to the issues raised above.

Chapter 2, Multi-Armed Bandits We first review the basic results of the classical multi-armed bandit problem. Some of them will be useful later in the combinatorial context.

Chapter 3, General Framework for Semi-Bandit Feedback We formalize here the framework of combinatorial stochastic semi-bandits with probabilistically triggered arms (which, we recalled, encompasses the usual framework). We then give a multitude of applications of this framework that can be found in the literature. Finally, we provide general technical results that will be useful throughout the thesis to prove upper bounds on policy regrets. More precisely, the theorems stated depend on the type of error we want to control (which is nothing else than the type of bonus we use). For the ℓ_∞ analysis (i.e., based on an ℓ_∞ type confidence region, or in an equivalent way on an ℓ_1 type bonus), we have a theorem that slightly improves (and also simplifies) the ℓ_∞ analysis of Wang and Chen (2017). As for the ℓ_2 analysis, we provide several new results, extending the work of Degenne and Perchet (2016b), and surpassing the ℓ_∞ analysis (gaining a factor m, up to a polylogarithmic factor). In doing so, we manage to associate the ℓ_2 analysis with the probabilistically triggered arms framework. This chapter contains some unpublished results, and some improvements of results published in:

- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020a) "Budgeted Online Influence Maximization", in 37th International Conference on Machine Learning (ICML).
- Pierre Perrault, Vianney Perchet, Michal Valko (2020) "Covariance-adapting algorithm for semi-bandits with application to sparse rewards", in 33rd Conference on Learning Theory (COLT).

- Pierre Perrault, Etienne Boursier, Vianney Perchet, Michal Valko (2020) "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits", in 34th Conference on Neural Information Processing Systems (NeurIPS).

Chapter 4, An Example of CMAB-T Problem: Sequential Search-and-Stop Here we present a new example of a problem that falls within both the probabilistically triggered arms and the budget frameworks. We call this problem *sequential search-and-stop*. It is described as follows: we consider a directed acyclic graph, where each node is an arm associated with a cost. There is an object that is randomly hidden in one of the nodes (according to an unknown distribution to be learned). In a round, the agent can inspect nodes, chosen one by one, with the constraint that a node is accessible if its in-neighbors have already been inspected. It can decide at any time to stop its search, and thus move to a new independent instance (where the object is randomly placed again). The agent has an initial budget, and each time a node is inspected, its price is paid. The objective of the agent is to find a maximum number of objects (in expectation). We provide a policy based on an ℓ_∞ approach, describing in particular a quasi-optimal and efficient strategy for the "offline" problem (i.e., when the distribution is known to the agent, who must then focus on exploitation). This chapter is adapted from the following publication:

- Pierre Perrault, Vianney Perchet, Michal Valko (2019b) "Finding the bandit in a graph: Sequential search-and-stop", in 22nd International Conference on Artificial Intelligence and Statistics (AIStats).

Chapter 5, The Structure of Uncertainty We begin by recalling the two main methods for obtaining ℓ_2 confidence regions: the *Laplace method*, and the *covering argument*. We then look more deeply into the structure of the ℓ_2 bonuses, and discover that as functions on $\mathcal{P}([n])$, they are submodular. We then use this structure to propose several approximation algorithms to maximize the sum of a linear function ($A \mapsto \mathbf{e}_A^\intercal \overline{\boldsymbol{\mu}}_{t-1}$) and a submodular function (the exploration bonus). This allows us to obtain approximations of ESCB for *matroid* constraints (a certain type of action space $\mathcal{A}$ generalizing the constraint on the cardinality). These approximations are computationally efficient and do not degrade the regret rate. We also provide an extension of this method to the budgeted framework. This chapter is based on the following publication:

- Pierre Perrault, Vianney Perchet, Michal Valko (2019a) "Exploiting structure of uncertainty for efficient matroid semi-bandits", in 36th International Conference on Machine Learning (ICML).

Chapter 6, Budgeted Online Influence Maximization The purpose of this chapter is to further investigate another well-known example of a problem involving probabilistically triggered arms, namely online influence maximization. As a reminder, in this problem, an agent actively learn a social network by interacting with it repeatedly, trying to find the best sets of influencers. We incorporate a budgeted framework to this problem, considering the total cost of an advertising campaign instead of the (per round) cardinality constraint. Our approach thus better models the real-world context where the cost of influencers varies and where advertisers want to find the best value for their overall budget. We propose ℓ_∞ and ℓ_2 policies, using the submodularity of an upper bound on ℓ_2 bonuses. This chapter is adapted from the following publication and prepublication:

- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020a) "Budgeted Online Influence Maximization", in 37th International Conference on Machine Learning (ICML).
- Pierre Perrault, Jennifer Healey, Zheng Wen, Michal Valko (2020b) "On the Approximation Relationship between Optimizing Ratio of Sub-modular (RS) and Difference of Submodular (DS) Functions", submitted.

Chapter 7, Covariance-Adapting Policy The goal here is to offer a more practical, and more accurate solution to build an ℓ_2 confidence region. Indeed, the ones we have seen so far are either based on mutual independence of outcomes, or due to the unrealistic knowledge of a sub-Gaussianity matrix $\mathbf{\Gamma}$. We consider a new general family of *sub-exponential* distributions parameterized by the covariance matrix, assumed to be unknown. This family contains the bounded support distributions and the Gaussian distributions. We prove a new lower bound on the regret of this family, which is parameterized by the covariance matrix, instead of the sub-Gaussianity matrix. We then construct an algorithm that uses covariance estimates, and provide an asymptotic analysis of the regret, reaching the lower bound. Finally, we apply and extend our results to the family of *sparse* outcomes, which has applications in many recommender systems. This chapter is adapted from the following publication:

- Pierre Perrault, Vianney Perchet, Michal Valko (2020) "Covariance-adapting algorithm for semi-bandits with application to sparse rewards", in 33rd Conference on Learning Theory (COLT).

Chapter 8, Statistical and Computational Efficiency of Thompson Sampling In this chapter, we consider an interesting alternative to the optimistic methods considered so far, namely *Thompson sampling*. In addition to its empirical superiority (Chapelle and Li, 2011), Thompson sampling is interesting in the context of combinatorial bandits because the action to be played is easy to compute: one should consider neither bonus nor maximum on a confidence region, because the parameter to be used in the objective function is simply taken at random, according to a well-chosen prior. Thompson sampling could thus answer the question of the existence of a (computationally) efficient policy with optimal asymptotic regret (up to a polylogarithmic factor in m), which is still open for many families of distributions, including mutually independent outcomes, and more generally the family of multivariate sub-Gaussian outcomes. 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 bound for the regret of CTS, using a beta prior. We then examine the more general framework of multivariate sub-Gaussian outcomes and propose a tight bound for the regret of CTS using a Gaussian prior. This latter result gives us an efficient alternative to the ESCB policy. This chapter is adapted from the following publication:

- Pierre Perrault, Etienne Boursier, Vianney Perchet, Michal Valko (2020) "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits", in 34th Conference on Neural Information Processing Systems (NeurIPS).

Chapter 9, Conclusion and Perspectives This chapter draws conclusions from the thesis and provides some directions for future work.

Chapter 2

Multi-Armed Bandits

This chapter is an introduction to the theory of *stochastic multi-armed bandits* (MAB), that is the simplest setting of reinforcement learning (Sutton and Barto, 1998). It is organized as follows: we first formally introduce the MAB problem. Next, we present different examples of motivation. Then we provide the basic theoretical results and introduce some MAB policies by giving guarantees on their performance. At the end of this chapter, we examine possible extensions of the classical MAB problem that are of particular interest in this thesis.

2.1 The stochastic multi-armed bandits problem

We consider a statistical model with incomplete observations. An automated *agent* partially observes a certain random phenomenon, described by a random vector $\mathbf{X} \in \mathbb{R}^n$, for a fixed integer $n \in \mathbb{N}^*$. We assume for the moment that the agent has no prior information on the distribution $\mathbb{P}_{\mathbf{X}}$, except that the mean $\boldsymbol{\mu}^* \triangleq \mathbb{E}[X]$ exists but is unknown. As usual in statistics, the phenomenon is accessible through a random process $(\mathbf{X}_t) \overset{iid}{\sim} \mathbb{P}_{\mathbf{X}}$. The observations are incomplete in the sense that the vector $\mathbf{X}_t$ is never observed entirely by the agent at any round t, but only one of its components, $X_{i_t,t}$, where the index i_t is selected by the agent at the beginning of round t. The quantity $X_{i_t,t}$ also represents a *gain*, or *reward* for the agent.

The objective is not directly to correctly estimate the distribution $\mathbb{P}_{\mathbf{X}}$ (so it is not really a statistical problem in the traditional sense), but rather to maximize the expected cumulative reward

$$\mathbb{E}\left[\sum_{t=1}^{T} X_{i_t,t}\right]$$

over the sequence of index choices (i_t), for a fixed *unknown* time horizon T. This is a learning problem called *stochastic multi-armed bandits* (MAB), which was first introduced in Robbins (1952). It takes its name from a casino slot machine. Indeed, it is worth remembering that a synonym for "slot machine" is "one-armed bandit": a slot machine has only one arm. Another view on the MAB problem is the following: we consider that the agent is facing n machines, or, more simply, an unusual machine, with n arms (numbered from 1 to n), which is why we talk about multi-armed bandits. Each of the n arms i, as soon as it is pulled, draws a reward from $\mathbb{P}_{X_i}$ at random. Distributions $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_n}$ are unknown and may be different from each other. In particular, some arm might be better for the agent, that is, they might return a higher reward. The machine is without memory: this ensures that the payment vectors $\mathbf{X}_t$ are effectively iid when t varies. Imagine now that the agent wants to play T times on the casino slot machines. The MAB problem raises the following natural question for the agent: what strategy to adopt for these T draws in order to maximize the expected total gain?

We will see that a smart allocation strategy for the agent has to trade-off between two behaviors:

- *Exploring* to collect information on the distribution $\mathbb{P}_{\mathbf{X}}$, and in particular, to estimate $\boldsymbol{\mu}^*$ effectively.
- *Exploiting* the arm that seems to be better given the rewards already observed.

Unlike the usual statistical problems, the agent is not interested in the proper description of the underlying stochastic phenomenon, but rather in the consequences of this good modeling in terms of reward.

Remark 1. *Although we use the notation $\mathbb{P}_{\mathbf{X}}$, we must keep in mind that an MAB problem is defined only by the marginals $\mathbb{P}_{X_1}, \dots, \mathbb{P}_{X_n}$. Indeed, we saw above tow ways of simulating the sequence $(i_1, X_{i_1,1}, i_2, X_{i_2,2}, \dots)$: at each round, the first uses $\mathbb{P}_{\mathbf{X}}$, whereas the second only uses a single marginal. As a consequence, for any distribution $\mathbb{P}_{\mathbf{X}}$, the two MAB instances defined by reward distributions $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ are the same. In particular, we can assume that $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ in this chapter. We'll see that this is no longer the case in stochastic combinatorial semi-bandits.*

2.1.1 The feedback

In our description of the MAB problem, the agent observes only the reward for the selected arm i_t, and nothing else. In particular, it does not observe rewards for other actions that could have been selected. This type of feedback is called *bandit feedback*. Another well-known type of feedback is called *full information*, where the whole vector $\mathbf{X}_t$ is observed by the agent at every round t (the observation is thus no longer partial), allowing it to focus on exploitation only. In this chapter, we will always assume that the agent has bandit feedback, that is a more realistic scenario (yet more challenging).

2.1.2 Mathematical definition of a policy

We mentioned the term "strategy" above, without really defining it formally. First of all, we will rather use the term *policy*, to emphasize that it gives the agent's way of behaving at a given time. We already have a pretty good intuition about what a policy is allowed to do, namely that the arm i_t is chosen in the round t based only on the information available at the beginning of the round, which is formed by past actions and gains. The arm i_t can also be chosen using an extra source of randomness. As a result, the agent chooses $i_t \in [n]$ according to a probability distribution $\mathbb{P}_{i_t}$ built with the past information. By universality of the standard uniform distribution, we can model the extra source of randomness needed to sample from $\mathbb{P}_{i_t}$ by the random variable U_t, where $(U_t) \overset{iid}{\sim} \mathcal{U}(0,1)$, independently from $(\mathbf{X}_t)$. Here is the same definition of a policy, but formulated in a rigorous mathematical way.

Definition 1 (Policy)**.** *A policy π is a sequence of random variables $\pi = (i_t) \in [n]^{\mathbb{N}^*}$ such that for all $t \geq 1$, i_t is $\mathcal{F}_t$-measurable, where $(\mathcal{F}_t)$ is the filtration defined by*

$$\mathcal{F}_t = \begin{cases} \sigma(U_1) & \text{if } t = 1 \\ \sigma(U_1, X_{i_1,1}, \dots, U_{t-1}, X_{i_{t-1},t-1}, U_t) & \text{if } t \geq 2. \end{cases}$$

When U_t is actually used in the choice of i_t, the policy is referred to as randomized.

Remark 2. *In addition to being measurable with respect to $\mathcal{F}_t$, the choice i_t can also depend on n and t. However, since T is unknown, i_t can't depend on it.*

2.1.3 A performance metric: the regret

The expected cumulative reward provides only an absolute assessment of the agent's total gain, but does not directly measure the relative deviation of this gain from the optimal total gain. From Definition 1, we can rewrite the expected cumulative reward of a policy $\pi = (i_t)$ as

$$\mathbb{E}\left[\sum_{t=1}^{T} X_{i_t,t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\,X_{i_t,t}|\mathcal{F}_t]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mu^*_{i_t}\right].$$

We see that the optimal policy for the MAB problem is $\pi^* \triangleq (i^*, i^*, \dots)$, where $i^* \in \arg\max_{i\in[n]} \mu^*_i$. In other words, when the distribution $\mathbb{P}_{\mathbf{X}}$ is known, the best choice is to pull at each round the arm with the largest mean. But since the agent first has to explore the environment to get information about the reward distribution of each arm, unavoidably some sub-optimal arms will be pulled. Thus, to evaluate the performance of a given policy π, we consider the *regret* of the agent for not playing optimally:

Definition 2 (Expected cumulative regret)**.** *The* expected cumulative regret *(that we shall simply call* regret*) for a policy $\pi = (i_t)$ is*

$$R_T(\pi) \triangleq T\mu^*_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} X_{i_t,t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{i_t}\right],$$

where for any $i \in [n]$, we define the gap *between means of arm i^* and arm i as $\Delta_i \triangleq \mu^*_{i^*} - \mu^*_i$.*

The goal for the agent is to play according to a policy π such that $R_T(\pi)$ is as small as possible.

Remark 3. *The regret (for a policy $\pi = (i_t)$) can be rewritten as $\pi = (i_t)$ is*

$$R_T(\pi) = \sum_{i\in[n]} \Delta_i \mathbb{E}[N_{i,T}],$$

where for any round $t \in [T]$, and any arm $i \in [n]$, $N_{i,t} \triangleq \sum_{t'=1}^{t} \mathbb{I}\{i_t = i\}$ is the counter of the number of times arm i was chosen by π during the t first rounds. This expression follows from Wald's identity.

2.1.4 A simple example: the Bernoulli bandit problem

As we will see in the next section, perhaps the simplest and most widely used reward distribution is the Bernoulli distribution, when the reward of each arm is either 0 or 1 ("failure or success", "heads or tails", etc.). This reward distribution is fully specified by the mean reward $\boldsymbol{\mu}^*$. We present in Figure 2.1 a very simple example of MAB problem with $n = 4$ arms and $T = 50$, with a naive policy π_{RAND} that selects the arm i_t uniformly at random.

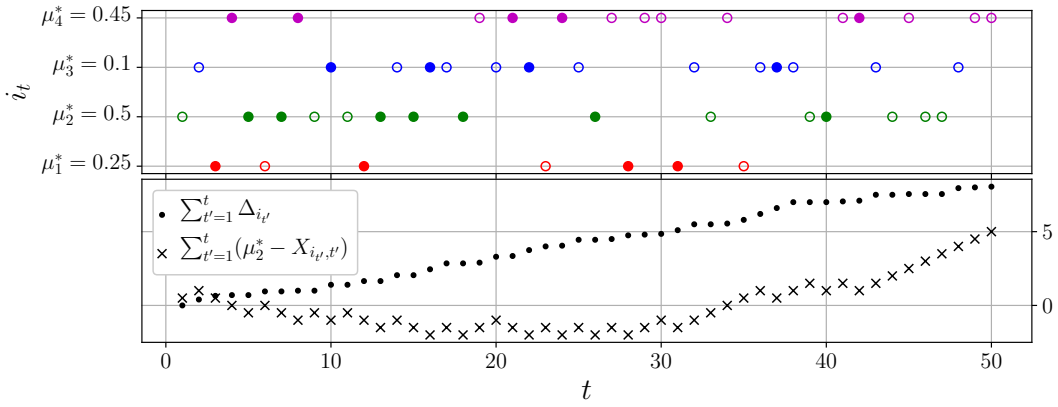

FIGURE 2.1: A Bernoulli bandit toy example. At the top, the points represent the choices i_t. When they are "full", the corresponding reward is 1, and when they are "empty", it is 0. At the bottom, we have plotted two types of cumulative regret curves that are both worth $R_t(\pi_{\text{RAND}})$ in expectation.

2.2 Some real world applications

This section presents some of the applications found in the MAB literature. This list (far from exhaustive) is intended to illustrate the great diversity of applications of the MAB framework and to motivate its study in this thesis.

2.2.1 Clinical trials

MAB models are historically guided by clinical trials (Thompson, 1933). Optimal clinical trial design is indeed a typical motivating application (Villar, Bowden, and Wason, 2015; Aziz, Kaufmann, and Riviere, 2019), although very little of the resulting theory has actually been used in clinical trial design and analysis. On the contrary, MAB models are now widely studied with completely different applications in mind (as we will see in the next subsections).

In a clinical trial, a set of n different drugs are tested to treat a disease. The most typical application is the choice between an old and a new drug (i.e., $n = 2$). It must be determined as soon as possible whether the new drug should be adopted or the old one maintained. Any error would result in lost human lives (or, at least in people suffering from disorders resulting from either incomplete treatment or excessive side effects). During the test phase, patients come sequentially to be administered one drug. It is assumed that the success of a drug on a patient is a random variable whose distribution depends on the drug in question. This variable can, for instance, be drawn from a Bernoulli distribution, and be worth 1 if the patient survives and 0 otherwise. The objective is then to maximize the total number of patients cured, or more precisely to achieve results that are almost as good as the best drug among the two. In particular, it is important to avoid using the sub-optimal drug too often, but rather to focus as soon as possible on the best one. Thus, a trade-off appears between collecting information about the two drugs to estimate their performance and using the information obtained so far to be efficient.

2.2.2 Web document ranking

MABs can also be used to rank web documents (Radlinski, Kleinberg, and Joachims, 2008) and more specifically web pages. The objective is to improve the performance of a Web search engine by increasing the probability that users will click on the highest-ranked results. The problem can be formalized as follows. We consider a fixed set of n documents and a fixed search query. Suppose we have a population of users, where each user has a personal set of documents considered as relevant with regard to the query, as well as a tolerance threshold giving the maximum number of documents the user is willing to check. Intuitively, users with different interpretations of the query have different relevant sets. At any round t, a (random) user comes hoping to find a document related to the query. An ordered set of m documents (among the n) is then proposed to the user. Results are consulted in order, and the goal is to avoid as much as possible the abandonment phenomenon, where no relevant document is presented before the user has reached its tolerance on the number of documents seen. The reward gained at round t is 1 if the user clicks on some document proposed, and 0 otherwise.

A first approach would be to consider each possible m-ordered set as an arm. However, Radlinski, Kleinberg, and Joachims (2008) rather proposed a more efficient method that runs m MAB instances in parallel. Each instance $i \in [m]$ has n arms and is in charge of choosing the document presented at rank i. If an MAB instance selects a document that has already been selected, the choice is saved, but another arbitrary unselected document is proposed instead. When the user clicks on document i, all the MAB instances that had selected the arm i as their first choice receive a reward of 1, and the others receive a reward of 0.

We will see in subsection 3.3.1 that this problem can be formulated as a stochastic combinatorial semi-bandit called "Cascading bandits".

2.2.3 Algorithm selection and black-box stochastic optimization

MABs have been used for algorithm selection in Gagliolo and Schmidhuber (2010). We give a description of the setting in the following. Let's consider a problem for which one can easily evaluate the quality of an algorithm to solve it. For example, the quality can be assessed by the processing time to deliver a solution, or by the value that an objective function takes at that solution. Suppose we have n algorithms (which may differ only by an extra parameter to tune), candidates to solve the problem, and that algorithms have some internal stochasticity, or that they can only be evaluated with additional noise (this is often the case for the running time). We can consider this setting as an MAB problem, where arms are the n algorithms, and rewards are (for instance) minus the time needed for the selected algorithm to run on the problem. MAB policies thus translate into strategies to automatically select the best algorithm among the n.

Web-based recommendation of products (films, music, articles, etc.) is an example where this method might be used. Indeed, in order to design a recommender system on their platform, companies collect enough data on their customers' preferences, as well as contextual metadata, and use collaborative or content-based filtering methods to deliver a recommendation. Now, if n recommender systems are designed in different ways, a company might want to know which one is the best to use, which is nothing more than an algorithm selection problem.

2.2.4 Online advertising

MABs represent a natural framework for expressing problems related to advertisement placement on a web page. This type of application has been widely and intensively studied over the last decade (Pandey and Olston, 2007; Babaioff, Sharma, and Slivkins, 2009; Li, Chu, et al., 2010; Li, Karatzoglou, and Gentile, 2016), stimulated by e-commerce and the need for many companies to provide personalized content. The problem can be described as follows. Given a fixed search query, there are n advertisers who wish to advertise on the search engine result for this query. The search engine wants to know which ad is the best to display, i.e., which one generates the most clicks from users. Specifically, when a user arrives hoping to find a web page related to the query, the search engine may display one of the n ads, hoping that the user will click on it. The objective in selecting ads is to maximize the total income of the search engine (each click on an advertisement proposed by the search engine yields a certain income).

It is sometimes simpler to break down the rounds not according to users (because the search engine cannot necessarily afford to change the advertising decision to each user, mainly for computing time reasons), but according to time (e.g., a round equals a day). Thus, for a given day t, the engine will choose the ad it will display throughout the day. A standard measure of the displayed ad performance at the end of the day is the *click-through rate* (CTR), defined as the ratio between the number of clicks and the number of users who saw the ad.

2.2.5 Cognitive radios

Opportunistic access to spectrum is the most common application context for cognitive radios. In telecommunications, certain frequency bands are reserved for certain uses, such as mobile communications, television or military communications. These uses are called primary uses and require a license. Although most of the spectrum that can be used for communications today is already allocated, it is left free most of the time by the primary users. It is to compensate for this under-use of spectrum that opportunistic access to spectrum has been considered. It allows secondary users to use licensed bands left free by primary users. Typically, a secondary user has a set of n frequency channels in which to transmit. Each of these channels is more or less used by primary users. The secondary user must then use the most vacant channel for its communications. At each round t, the secondary user chooses to transmit through one of the n channels, and receives a reward of 1 if the selected transmission channel is vacant (not used by a primary user, the transmission is effective), and 0 if that channel is busy (no communication is possible through the selected channel). This setting can be seen as an MAB problem when the use of a channel by primary users is uncertain, and the reward received can, therefore, be modeled as a Bernoulli random variable.

The applications of MABs to cognitive radios have been studied in a more general context of multiplayer bandits, where several secondary users want to use the telecommunication network simultaneously (Rosenski, Shamir, and Szlak, 2016; Besson and Kaufmann, 2017; Lugosi and Mehrabian, 2018; Boursier and Perchet, 2019). Opportunistic access to spectrum is a major issue for wireless technologies, in particular for 5G technology (Wang, Song, et al., 2019).

Algorithm 1 A MCTS algorithm

Initially, the tree has a single node (the root r, the current state of the game).
The following steps are repeated T times, and ultimately, the most visited child of r is returned.
Selection: From r, use successively the MAB algorithm assigned to the current node to select a next child node until a leaf ℓ is reached.
Expansion: Unless ℓ is terminal, create a child node c in the tree that is any valid move from the game position defined by ℓ.
Simulation: Complete a random play-out from c by choosing uniform random moves until the game is decided.
Back-propagation: Use the result of the play-out as a reward for each MAB algorithm on the path from c to r.

2.2.6 Monte-Carlo tree search

Monte Carlo tree search (MCTS) is a heuristic search algorithm used in some kind of decision processes. It is notably used in games (see Browne et al. (2012) and the references therein for different variants of MCTS and applications to games and other search, optimization, and control problems). We can mention its implementation in recent computer programs of Go, such as Crazy-Stone (Coulom, 2007), MoGo (Gelly et al., 2006; Wang and Gelly, 2007), AlphaGo (Silver, Huang, et al., 2016), and AlphaGoZero (Silver, Schrittwieser, et al., 2017). Given the current state of the game, the goal is to determine the next action to be selected by the player. MCTS algorithms explore the tree of possibilities in a non-symmetric way. The root is the current configuration of the game, and each node is a configuration and its children are the following possible configurations. The MAB setting has been used to guide exploration in the tree structure, allowing to perform efficient tree search by assigning an MAB algorithm to each node and following an optimistic search strategy that explores in priority the most promising branches (see Algorithm 1).

The intuition of Algorithm 1 is that at a given node there are some possible choices, i.e., arms, corresponding to the child nodes, and the use of an MAB algorithm should enable the selection of the best arm given noisy rewards samples. The reason why this algorithm has shown good performance in practice in several large tree search problems is that it prioritizes the most promising branches according to previously observed sample rewards. This is very useful in situations where the reward function has some smoothness property (i.e., the initial random reward samples provide information on where the search should focus). Theoretically, however, since the reward samples obtained from any node are not iid, the algorithm does not have nice finite-time performance guarantees (see Coquelin and Munos (2007)).

One of the best known algorithms extends the *Upper Confidence Bound* algorithm (UCB) of Auer, Cesa-Bianchi, and Fischer (2002) (which we will explore in more detail in section 2.4) to tree-structured search, defining the *Upper Confidence bound applied to Trees* algorithm (UCT) (Kocsis and Szepesvári, 2006).

2.2.7 Network exploration

Madhawa and Murata (2019) studied the problem of exploring a large unweighted and undirected graph $G = (V, E)$. This has applications in collecting information about user profiles and their friends in a social network G. In their setting, G cannot be fully observed, and only a subgraph $G'_t = (V'_t, E'_t)$ is available to the agent at time t. Initially, G'_0 is only composed of a single node. At each round t, the agent

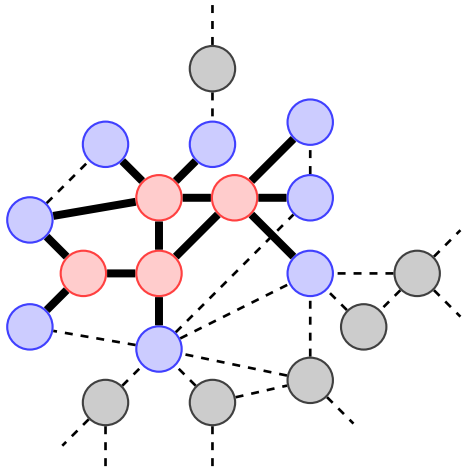

FIGURE 2.2: Example of an exploration graph G'_t. Red nodes are probed, blue nodes are only observed, grey nodes exist in the original graph G and are yet to be observed. Vertices of the graph G'_t are red and blue nodes, and edges of G'_t are represented with bold lines. Dash lines are used to denote unobserved links at the given moment t. To build G'_{t+1}, the agent has to choose a blue node. Then, it becomes red and its neighborhood in G becomes blue.

selects a node $i_t \in V'_t$ to probe in order to expand G'_t by adding the neighbors of i_t. More precisely, when i_t is selected, G'_{t+1} is obtained from G'_t as follows. $V'_{t+1} = V'_t \cup \mathcal{N}_{i_t}$, where $\mathcal{N}_{i_t} \triangleq \{j \in V, \{i_t, j\} \in E\}$ is the neighbourhood of i_t in G, and $E'_{t+1} = E'_t \cup \{\{i_t, j\}, j \in \mathcal{N}_{i_t}\}$. An example of the graph G'_t is given in Figure 2.2. The goal for the agent is to discover as many nodes as possible within a limited time budget T, i.e., $|V'_T|$ has to be maximized.

As for MCTS, MAB algorithms might be useful heuristics, although the problem watched is not strictly speaking an MAB problem. In order to do so, Madhawa and Murata (2019) considered that each node i is associated with a certain set of features F_i (e.g., interests, liked pages, etc.), and that the neighborhood size of i (which is a reward for the agent if it probes i) could be represented as a sample according to an unknown distribution P_{F_i} depending on F_i. In other words, nodes with similar features in the observed network will result in similar rewards. The trade-off between exploration and exploitation thus comes naturally. Indeed, for exploitation, the agent can consider choosing $i_t = i$ based on the a priori quality of F_i, which can be estimated by looking at the past performance of nodes j, with F_j close to F_i. For exploration, the criterion to be taken into account is the amount of information brought by a new reward sampled from P_{F_i}, or equivalently how much information the agent has already collected on features close to F_i.

2.3 Regret lower bounds

Before giving examples of MAB policies, it is important to know to what extent the agent can hope to minimize the expected cumulative regret. Indeed, we already saw that the agent, not knowing which arm is the best, must inevitably suffer some regret. A lower bound on the regret would, therefore, provide information on what the agent can aim for. However if one considers, as a performance measure, $R_T(\pi)$ for a single reward distribution $\mathbb{P}_{\mathbf{X}}$ without restricting the class of admissible policies π, then

it is clear that we can achieve $R_T(\pi) = 0$. Indeed, policies that always chose the same arm are admissible, and one of them has a 0 regret (it is obvious that these policies are of no interest because they do not learn, and although they are very good on a specific problem, they are very bad on the others). Thus, the regret must be considered either in more than one instance $\mathbb{P}_{\mathbf{X}}$ at the same time (Definition 3), or by limiting the class of policies allowed (Definition 4). This gives two kinds of regret lower bound.

Definition 3 (*Distribution-free* lower bound)**.** *A distribution-free lower bound is a lower bound on*

$$\inf_{\pi} \sup_{\mathbb{P}_{\mathbf{X}} \in \mathcal{D}} R_T(\pi),$$

where $\mathcal{D}$ is a family of reward distributions.

Definition 4 (*Distribution-dependent* lower bound)**.** *A distribution-dependent lower bound is a lower bound on $R_T(\pi)$ that holds for any distribution $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$ and any policy $\pi \in \Pi$, where Π is a family of admissible policies and $\mathcal{D}$ is a family of reward distributions.*

For distribution-free lower bounds, the performance is measured considering the worst possible reward distribution $\mathbb{P}_{\mathbf{X}}$ (specific to the policy and the horizon T). This kind of lower bound doesn't depend on $\mathbb{P}_{\mathbf{X}}$. On the other hand, distribution-dependent lower bounds hold for a fixed distribution $\mathbb{P}_{\mathbf{X}}$ and depend on it. Only a restricted class of policies Π is considered in order to avoid the consideration of policies of little interest that gives a small regret for the distribution $\mathbb{P}_{\mathbf{X}}$, but perform badly on other instances.

We will see in the following that these two measures may have very different behaviors in T.

2.3.1 Distribution-dependent lower bound

Lai and Robbins (1985) proved a lower bound of the expected cumulative regret of order $\log(T)$ in a particular parametric framework. This work has then been extended by Agrawal (1995) and Burnetas and Katehakis (1996). These papers deal with the class of *consistent policies*, defined as follows.

Definition 5 (Consistent policies)**.** *A consistent policy π (with respect to $\mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$) is such that for all $a > 0$ and all reward distribution $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$, the expected cumulative regret satisfies*

$$R_T(\pi) =_{T\to\infty} o(T^a).$$

Consistent policies offer good performance for all arm distribution in $\mathcal{D}$. The authors were interested in the price to be paid for a policy to be consistent. In other words, they ask the question *What is the minimum regret that a consistent policy must suffer?* The logarithmic bound of Agrawal (1995) and Burnetas and Katehakis (1996) answers this question. It uses the Kullback–Leibler divergence (KL) (Kullback and Leibler, 1951), defined in Definition 6.

Definition 6 (Kullback–Leibler divergence (KL))**.** *The Kullback–Leibler divergence between two probability distributions P and Q is defined as*

$$\mathrm{KL}(P\|Q) \triangleq \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP, & \text{if } P \text{ is absolutely continuous with respect to } Q \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{dP}{dQ}$ *is the Radon–Nikodym derivative of* P *with respect to* Q.

The KL is a measure of how one probability distribution is different from a second one. Intuitively, it measures the rate (in terms of the number of samples) at which the two distributions can be distinguished based on the samples from the first distribution. The smaller the KL divergence, the more difficult it is to distinguish the two distributions. The KL divergence is non-negative, and the value zero is reached if and only if the distributions are identical. it is generally asymmetric in the two distributions.

We are now ready to state the lower bound of Agrawal (1995) and Burnetas and Katehakis (1996), in Theorem 1.

Theorem 1 (Agrawal (1995) and Burnetas and Katehakis (1996))**.** *Let* π *be a consistent policy with respect to a distribution family* $\mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$. *Then we have, for all* $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$,

$$R_T(\pi) \geq \sum_{i \in [n],\ \Delta_i > 0} \frac{\log(T)(1 - o(1))\Delta_i}{\inf_{P \in \mathcal{D}_i(\mu_{i*}^*)} \mathrm{KL}(\mathbb{P}_{X_i} \| P)},$$

where $\mathcal{D}_i(\mu) \triangleq \left\{ \mathbb{P}_{X_i'} \in \mathcal{D}_i,\ \mathbb{E}[X_i'] > \mu \right\}$.

Proof. We fix $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$ and $i \in [n]$ such that $\Delta_i > 0$. In the proof, we use the notation $d_i = \inf_{P \in \mathcal{D}_i(\mu_{i*}^*)} \mathrm{KL}(\mathbb{P}_{X_i} \| P)$. To get the desired result, it is sufficient to prove the following lower bound (and then sum over i, weighting by Δ_i)

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_{i,T}]}{\log(T)} \geq {d_i}^{-1}.$$

Notice that it is equivalent to prove that

$$\forall \varepsilon \in (0,1),\ \liminf_{T \to \infty} \frac{\mathbb{E}[N_{i,T}]}{\log(T)} \geq (1 - \varepsilon){d_i}^{-1}.$$

We thus fix $\varepsilon \in (0,1)$. From Markov's inequality we get

$$\frac{\mathbb{E}[N_{i,T}]}{\log(T)} \geq \frac{(1-\varepsilon)\mathbb{P}\left[N_{i,T}/\log(T) \geq (1-\varepsilon){d_i}^{-1}\right]}{d_i}.$$

Thus, it suffices to show that

$$\lim_{T \to \infty} \mathbb{P}\left[\frac{N_{i,T}}{\log(T)} \geq \frac{1-\varepsilon}{d_i}\right] = 1,$$

or equivalently that

$$\lim_{T \to \infty} \mathbb{P}\left[\frac{N_{i,T}}{\log(T)} < \frac{1-\varepsilon}{d_i}\right] = 0. \tag{2.1}$$

Let $\delta > 0$ be such that $(1-\delta)/(1+\delta) > 1 - \varepsilon$. By definition of d_i, there exists $P \in \mathcal{D}_i(\mu_{i*}^*)$ such that

$$d_i < \mathrm{KL}(\mathbb{P}_{X_i} \| P) < (1+\delta)d_i. \tag{2.2}$$

Define the events

$$\mathfrak{A}_T \triangleq \left\{ \frac{N_{i,T}}{\log(T)} < \frac{1-\delta}{\mathrm{KL}(\mathbb{P}_{X_i} \| P)} \right\}, \quad \mathfrak{C}_T \triangleq \left\{ L_{N_{i,T}} \leq (1-\delta/2)\log(T) \right\},$$

where

$$L_t \triangleq \sum_{t'=1}^{t} \log\left(\frac{d\mathbb{P}_{X_i}}{dP}(X_{i,t'}) \right).$$

We write $\mathbb{P}', \mathbb{E}'$ the probability and expectation when samples from arm i are drawn from P instead of $\mathbb{P}_{X_i}$ (i is thus the best arm in this new environment). Notice, since $\mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$, and $P \in \mathcal{D}_i$, we have that

$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_{i-1}} \otimes P \otimes \mathbb{P}_{X_{i+1}} \otimes \cdots \otimes \mathbb{P}_{X_n} \overset{(2.3)}{\in} \mathcal{D}.$$

We now show that $\mathbb{P}[\mathfrak{A}_T] = \mathbb{P}[\mathfrak{A}_T \cap \mathfrak{C}_T] + \mathbb{P}[\mathfrak{A}_T \backslash \mathfrak{C}_T] \to_{T\to\infty} 0$, which together with (2.2) proves (2.1). On the one hand, we have

$$\begin{aligned}
\mathbb{P}[\mathfrak{A}_T \cap \mathfrak{C}_T] &\leq e^{(1-\delta/2)\log(T)} \mathbb{P}'[\mathfrak{A}_T \cap \mathfrak{C}_T] && (2.4)\\
&\leq T^{1-\delta/2} \mathbb{P}'[\mathfrak{A}_T] \\
&= T^{1-\delta/2} \mathbb{P}'\left[T - N_{i,T} > T - \frac{1-\delta}{\mathrm{KL}(\mathbb{P}_{X_i}\|P)} \log(T) \right] \\
&\leq \frac{T^{1-\delta/2} \mathbb{E}'[T - N_{i,T}]}{T - \frac{1-\delta}{\mathrm{KL}(\mathbb{P}_{X_i}\|P)} \log(T)} && (2.5)\\
&= \frac{T^{-\delta/2} \sum_{j\neq i} \mathbb{E}'[N_{j,T}]}{1 - \frac{1-\delta}{\mathrm{KL}(\mathbb{P}_{X_i}\|P)} \log(T)/T} \to_{T\to\infty} 0. && (2.6)
\end{aligned}$$

(2.4) is from a change of measure: the random variable $\mathbb{I}\{\mathfrak{C}_T \cap \{N_{i,T} = t\}\}$ is measurable with respect to $X_{i,1}, \ldots, X_{i,t}$ and $X_{j,1}, \ldots, X_{j,T}$ ($j \neq i$) (and a possible extra source of randomness). Its partial integral against $\prod_{t'=1}^{t} d\mathbb{P}_{X_i}(x_{i,t'})$ is lower than the one against $e^{(1-\delta/2)\log(T)} \prod_{t'=1}^{t} dP(x_{i,t'})$, by virtue of the event $\mathfrak{C}_T$. (2.5) is a consequence of Markov's inequality, and the limit in (2.6) is by consistency of π and by (2.3).

On the other hand, letting $b_T \triangleq \log(T)(1-\delta)/\mathrm{KL}(\mathbb{P}_{X_i}\|P)$, we have

$$\begin{aligned}
\mathbb{P}[\mathfrak{A}_T \backslash \mathfrak{C}_T] &\leq \mathbb{P}\left[\max_{t<b_T} L_t > (1-\delta/2)\log(T) \right] \\
&= \mathbb{P}\left[b_T^{-1} \max_{t<b_T} L_t > \frac{1-\delta/2}{1-\delta} \mathrm{KL}(\mathbb{P}_{X_i}\|P) \right].
\end{aligned}$$

This last term tends to zero, as a consequence of the law of large numbers. □

Example 1 (The Bernoulli bandit case)**.** *As we already saw in the previous section, a well studied particular case of MAB setting is where $\mathcal{D}$ is the family of distributions $\mathbb{P}_{\mathbf{X}}$ such that $\mathbf{X} \in \{0,1\}^n$, i.e, X_i follows a Bernoulli distribution for all $i \in [n]$. In this case, Theorem 1 is reformulated as follows: Let π be a consistent policy with respect to a distribution family $\mathcal{D}$. Then we have, for all $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$,*

$$R_T(\pi) \geq \sum_{i\in[n],\ \Delta_i>0} \frac{\log(T)(1-o(1))\Delta_i}{\mathrm{kl}(\mu_i^*, \mu_{i^*}^*)},$$

where $\mathrm{kl}(x,y) \triangleq \mathrm{KL}(\mathrm{Bernoulli}(x)\|\mathrm{Bernoulli}(y)) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$, $x, y \in [0,1]$.

In Example 1, the result can be interpreted as follows: To achieve good performance for all arm distributions, each sub-optimal arm i needs to be explored, asymptotically, no fewer than $\log(T)/\mathrm{kl}(\mu_i^*, \mu_{i^*}^*)$ times, which grows in a logarithmic order with T and is inversely proportional to the distribution divergence $\mathrm{kl}(\mu_i^*, \mu_{i^*}^*)$ between this sub-optimal arm i and the optimal arm i^*. Another classical example is given by Gaussian distributions.

Example 2 (The Gaussian bandit case)**.** *Consider the MAB setting where $\mathcal{D}$ is the family of distributions such that X_i follows a Gaussian distribution of variance σ_i^2 for all $i \in [n]$. In this case, Theorem 1 is reformulated as follows: Let π be a consistent policy with respect to a distribution family $\mathcal{D}$. Then we have, for all $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$,*

$$R_T(\pi) \geq \sum_{i\in[n],\ \Delta_i>0} \frac{2\sigma_i^2 \log(T)(1-o(1))}{\Delta_i}.$$

In case the distribution family $\mathcal{D}$ has not a product form, one can still obtain lower bounds, where the change of distribution argument is done in several arms instead of one. The important thing is to change the distribution while remaining within the family under consideration. We give in the following an example of such lower bound result in the Bernoulli bandit problem.

Theorem 2 (Graves and Lai (1997))**.** *Let π be a consistent policy with respect to a distribution family $\mathcal{D}$ such that for all $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$, $\mathbf{X} \in \{0,1\}^n$. Then we have, for all $\mathbb{P}_{\mathbf{X}} \in \mathcal{D}$,*

$$\liminf_{T\to\infty} \frac{R_T(\pi)}{\log(T)} \geq \inf_{\mathbf{c}\in\mathcal{C}} \sum_{i\in[n],\ \Delta_i>0} c_i\Delta_i,$$

where

$$\mathcal{C} \triangleq \left\{(c_i)_i \in \mathbb{R}_+^n,\ \forall \mathbb{P}_{\mathbf{X}'} \in \mathcal{B}(\boldsymbol{\mu}^*),\ \sum_i c_i \mathrm{kl}(\mu_i^*, \mathbb{E}[X_i']) \geq 1\right\},$$

$$\mathcal{B}(\boldsymbol{\mu}^*) \triangleq \left\{\mathbb{P}_{\mathbf{X}'} \in \mathcal{D},\ s.t.\ \max_i \mathbb{E}[X_i'] > \mu_{i^*}^*\ \textit{and}\ \mathbb{E}[X_{i^*}'] = \mu_{i^*}^*\right\}.$$

2.3.2 Distribution-free lower bound

We are now stating a distribution-free lower bound for the expected cumulative regret, in Theorem 3. This result dates back to Vogel (1960), and was more recently improved by Bubeck and Cesa-Bianchi (2012).

Theorem 3 (Bubeck and Cesa-Bianchi (2012))**.** *Assume that $n \geq 2$. Let $\mathcal{D}$ be the family of distributions $\mathbb{P}_{\mathbf{X}}$ such that $\mathbb{P}_{X_i}$ is a Bernoulli distribution for all $i \in [n]$. Then we have*

$$\inf_{\pi} \sup_{\mathbb{P}_{\mathbf{X}}\in\mathcal{D}} R_T(\pi) \geq \frac{1}{20}\sqrt{nT}.$$

Proof. For $i \in \{0,\dots,n\}$, we consider the bandit problem $P_i \in \mathcal{D}$ where all arm rewards are sampled from a Bernoulli distribution of mean $1/2$, except for one arm i that follows a Bernoulli distribution of mean $1/2+\varepsilon$, for $\varepsilon > 0$ to be stated later (thus, for $i = 0$, all arms have the same distribution). We denote (respectively)

$\mathbb{P}^{(i)}$, $\mathbb{E}^{(i)}$ and $R_T^{(i)}$ the probability, the expectation, and the regret when rewards are sampled from the distribution P_i. We have for all policy π

$$\begin{aligned}\sup_{\mathbb{P}_{\mathbf{X}}\in\mathcal{D}} R_T(\pi) &\geq \sup_{i\in[n]} R_T^{(i)}(\pi)\\ &\geq \frac{1}{n}\sum_{i\in[n]} R_T^{(i)}(\pi)\\ &= \frac{\varepsilon}{n}\sum_{i\in[n]}\Big(T - \mathbb{E}^{(i)}[N_{i,T}]\Big). \end{aligned} \tag{2.7}$$

Notice that

$$\inf_{\pi}\frac{1}{n}\sum_{i\in[n]} R_T^{(i)}(\pi) = \inf_{\pi \text{ deterministic}}\frac{1}{n}\sum_{i\in[n]} R_T^{(i)}(\pi),$$

since a randomized policy is a convex combination of deterministic ones. We thus assume that π is deterministic in the following. We can bound the difference between $\mathbb{E}^{(i)}[N_{i,T}] - \mathbb{E}^{(0)}[N_{i,T}]$ for all $i \in [n]$ as

$$\begin{aligned}\mathbb{E}^{(i)}[N_{i,T}] - \mathbb{E}^{(0)}[N_{i,T}] &= \sum_{t=1}^{T}\Big(\mathbb{P}^{(i)}[i_t = i] - \mathbb{P}^{(0)}[i_t = i]\Big)\\ &\leq T\Big\|\mathbb{P}^{(i)} - \mathbb{P}^{(0)}\Big\|_{\mathrm{TV}}\\ &\leq T\sqrt{\mathrm{KL}\Big(\mathbb{P}^{(0)}\|\mathbb{P}^{(i)}\Big)/2},\end{aligned}$$

where, for two probability distributions P and Q on a measurable space $(\Omega, \mathcal{A})$, $\|P - Q\|_{\mathrm{TV}} \triangleq \sup_{A\in\mathcal{A}}|P(A) - Q(A)|$ is the total variation distance between P and Q, and where the last inequality is from Pinsker's inequality (see Proposition 1). Since π is a deterministic policy, we get from the chain rule for the KL that

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}^{(0)}\|\mathbb{P}^{(i)}\Big) &= \mathbb{E}^{(0)}[N_{i,T}]\mathrm{KL}(P_0\|P_i)\\ &= \mathbb{E}^{(0)}[N_{i,T}]\mathrm{KL}(\mathrm{Bernoulli}(1/2)\|\mathrm{Bernoulli}(1/2+\varepsilon))\\ &= \mathbb{E}^{(0)}[N_{i,T}]\log(1/(1-4\varepsilon^2))/2\end{aligned}$$

Substituting the obtained upper bound on $\mathbb{E}^{(i)}[N_{i,T}]$ in (2.7) gives

$$\begin{aligned}\sup_{\mathbb{P}_{\mathbf{X}}\in\mathcal{D}} R_T(\pi) &\geq \varepsilon T - \frac{\varepsilon}{n}\sum_{i\in[n]}\mathbb{E}^{(0)}[N_{i,T}] - \frac{\varepsilon T}{2n}\sqrt{\log\Big(\frac{1}{1-4\varepsilon^2}\Big)}\sum_{i\in[n]}\sqrt{\mathbb{E}^{(0)}[N_{i,T}]}\\ &\geq \varepsilon T - \frac{\varepsilon}{n}T - \frac{\varepsilon T}{2n}\sqrt{\log\Big(\frac{1}{1-4\varepsilon^2}\Big)}\sqrt{nT},\end{aligned}$$

where this last inequality uses $\sum_{i\in[n]}\mathbb{E}^{(0)}[N_{i,T}] = T$ and the Cauchy–Schwarz type inequality $\sum_{i\in[n]}\sqrt{\mathbb{E}^{(0)}[N_{i,T}]} \leq \sqrt{n\sum_{i\in[n]}\mathbb{E}^{(0)}[N_{i,T}]}$. Setting $\varepsilon = \sqrt{n/T}/4$ and using $-\log(1-y) \leq 4\log(4/3)y$ for $y \in [0, 1/4]$ gives the final result. □

Proposition 1 (Pinsker's inequality, see e.g. Tsybakov (2009))**.** *Let P and Q be two distributions over some measurable space $(\Omega, \mathcal{A})$, then*

$$\mathrm{KL}(P\|Q) \geq 2\|P - Q\|_{\mathrm{TV}}^2.$$

Proof. Without loss of generality, we can assume that P is absolutely continuous with respect to Q, since otherwise $\mathrm{KL}(P\|Q) = \infty$. Recall that $\|P-Q\|_{\mathrm{TV}} = \sup_{A\in\mathcal{A}}|P(A)-Q(A)| = \frac{1}{2}\int\left|\frac{dP}{dQ}-1\right|dQ$. We thus have, with $r = dP/dQ - 1$, $f = |r|/g$ and $g = \sqrt{1+r/3}$,

$$\begin{aligned}
2\|P-Q\|_{\mathrm{TV}}^2 &= \frac{1}{2}\left(\int\left|\frac{dP}{dQ}-1\right|dQ\right)^2 \\
&= \frac{1}{2}\left(\int fg\; dQ\right)^2 \\
&\le \frac{1}{2}\left(\int f^2\; dQ\right)\left(\int g^2\; dQ\right) \qquad (2.8)\\
&= \frac{1}{2}\int f^2\; dQ \\
&= \int \frac{r^2}{2(1+r/3)}dQ \\
&\le \int ((1+r)\log(1+r)-r)\; dQ = \mathrm{KL}(P\|Q), \qquad (2.9)
\end{aligned}$$

where (2.8) is from Cauchy–Schwarz inequality, and (2.9) is from

$$0 \le h(x) = \frac{-x^2}{2(1+x/3)} + (1+x)\log(1+x) - x \quad \forall x \ge -1,$$

indeed, $h'(x) = \log(1+x) - 3x(x+6)/(2(x+3)^2)$ and $h''(x) = x^2(x+9)/((x+3)^3(x+1)) \ge 0$ (and equals 0 only for $x=0$), so h' is increasing and $h'(x) \ge h'(0) = 0$ if and only if $x \ge 0$, so $\min_{x\ge -1} h(x) = h(0) = 0$. □

The minimax nature of the formulation implies that the lower bound in Theorem 3 holds for all distribution families that include those with $\{0,1\}$ support, for example distributions with bounded support on $[0,1]^n$ or more generally distributions whose marginals are $1/4$-sub-Gaussian (see Definition 7 for the definition of a κ^2-sub-Gaussian distribution, and Proposition 2 for the fact that a random variable in $[0,1]$ is $1/4$-sub-Gaussian). The question is in the achievability of this lower bound for these more general sets of distributions.

Definition 7 (κ^2-sub-Gaussian distribution, Buldygin and Kozachenko (1980))**.** *We say that a random variable $X \in \mathbb{R}$ (or a distribution $\mathbb{P}_X$) is κ^2-sub-Gaussian if*

$$\forall \lambda \in \mathbb{R},\ \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le e^{\lambda^2\kappa^2/2}.$$

Proposition 2 (Hoeffding's Lemma, Hoeffding (1963))**.** *Let X be a random variable almost surely in $[a,b]$, with $a < b \in \mathbb{R}$. Then, X is $(b-a)^2/4$-sub-Gaussian.*

Proof. Let $\lambda \in \mathbb{R}$. We can assume that $\mathbb{E}[X] = 0$ by replacing a by $a - \mathbb{E}[X]$ and b by $b - \mathbb{E}[X]$. By convexity of the function $x \mapsto e^{\lambda x}$, we have for all $x \in [a,b]$,

$$e^{\lambda x} \le \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}.$$

Evaluating this inequality in X and taking the expectation, we get

$$\mathbb{E}\left[e^{\lambda X}\right] \le e^{\psi(\lambda(b-a))} = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b},$$

with

$$\psi(u) = -pu + \log(1 - p + pe^u), \quad p = \frac{-a}{b-a}.$$

We see that $\psi(0) = \psi'(0) = 0$. In addition,

$$\psi''(u) = \frac{(1-p)pe^u}{(1-p+pe^u)^2} \leq \frac{1}{4}.$$

Finally, thank to Taylor's theorem, there is $\theta \in (0,1)$ such that

$$\psi(u) = \psi(0) + \psi'(0)u + \psi''(\theta u)\frac{u^2}{2} \leq \frac{u^2}{8}.$$

□

Remark 4. *In Theorem 3, when instead $\mathcal{D}$ is the family of distributions $\mathbb{P}_{\mathbf{X}}$ such that $\mathbf{X}$ is in $\{a,b\}^n$ almost surely, for $a < b \in \mathbb{R}$, we have the lower bound*

$$\inf_{\pi} \sup_{\mathbb{P}_{\mathbf{X}} \in \mathcal{D}} R_T(\pi) \geq \frac{b-a}{20}\sqrt{nT}.$$

Indeed, we can trivially reduce to the case $\mathbf{X} \overset{a.s.}{\in} \{0,1\}^n$ by translating and scaling up all rewards X_i as

$$\frac{X_i - a}{b-a} \overset{a.s.}{\in} \{0,1\}.$$

This simply has the effect of dividing the regret by $b-a$, and can be seen as a simple convention taken on the unit of the rewards.

2.4 Policies

In this section, we review some examples of well known MAB policies. We will see in particular that the proposed policies have one thing in common, namely that they all make the decision i_t by maximizing a certain function defined on $[n]$ that depends on the agent observation history up to the round $t-1$. We refer to this kind of policy as *index-based* (the function value is the index that guides the decision).

2.4.1 Index-based MAB policies

A large majority of existing MAB policies are based on a scoring index maintained across rounds, which gives for each arm i and each round t a numerical value representing the potential quality of arm i after $t-1$ rounds. The score of the arm i can be computed from n, t, the history and a possible source of randomness, i.e., is measurable with respect to $\mathcal{F}_t$. These policies are described in the Algorithm 2 and work as follows. During the first n rounds, they sequentially play arms $1, 2, ..., n$ to perform the initialization. In all subsequent rounds, they compute for each arm the corresponding score and select the arm with the greatest one (ties are broken at random). The scoring index of arm i at round t can generally be seen as a proxy or estimates of the true mean μ_i^*. Indeed, ultimately, when the agent have collected enough knowledge to know the mean vector $\boldsymbol{\mu}^*$, the policy should behave as the optimal policy π^*, that is nothing else than Algorithm 2 with μ_i^* as the score of arm i. This is the reason why we sometimes use the notation $\mu_{i,t}$ for the score of arm i at round t.

Algorithm 2 Generic index-based MAB policy

Initialization: Play each arm once.
for $t \in \{n+1, \dots, T\}$ **do**
  For all $i \in [n]$, compute $\mu_{i,t}$, the score of arm i at round t.
  Play the arm $i_t \triangleq \arg\max_{i \in [n]} \mu_{i,t}$, ties are broken at random.
end for

2.4.2 Examples

We give in Table 2.1 some well-known examples of index-based MAB policies. The indexes provided are valid when the reward vector $\mathbf{X}$ is in $[0,1]^n$ almost surely (for the sake of simplicity, we limit ourselves to this case here). There are several quantities that we have to define in order to be able to read Table 2.1. First, the agent can estimate the mean μ_i^* of every arm i with their corresponding *empirical averages* defined as

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

for $t \geq 1$, where $N_{i,t-1} \triangleq \sum_{t' \in [t-1]} \mathbb{I}\{i = i_{t'}\}$ is the number of time arm i have been drawn for the first $t-1$ rounds, as mentioned above in Remark 3. Another quantity that can be used is the *empirical variance* of an arm i, defined as

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

for $t \geq 1$. Notice that another expression for the empirical variance (that is more practical to maintain for a policy) is

$$\overline{\sigma}^2_{i,t-1} = \frac{\sum_{t' \in [t-1]} \mathbb{I}\{i = i_{t'}\} X^2_{i,t'}}{N_{i,t-1}} - \overline{\mu}^2_{i,t-1}.$$

The advantage of maintaining such a quantity is that it allows the agent to measure how uncertain the rewards obtained are, and can therefore be useful in the control of the estimation error. For example, when the rewards of an arm i are deterministic, the corresponding empirical variance is zero, allowing the agent to limit its exploration on this arm.

We begin by describing ε-GREEDY (see Watkins (1989)), that is probably one of the simplest MAB policy. Then, we describe policies which can be qualified as optimistic, since they rely on the *optimism in face of uncertainty principle* (OFU principle). We then provide a Bayesian randomized policy, THOMPSON SAMPLING (TS), that maintain an estimated payoff distribution for each arm. Finally, we notice that these previous policies are rather destined to satisfy the obtimality criterion given in Theorem 1, and give MOSS, an example of a policy satisfying the second criterion of optimality, given by Theorem 3.

ε-greedy

ε-GREEDY is easy to understand, since it simply chooses the arm uniformly at random with probability ϵ, and choose the leader arm (the one with the highest empirical average) with probability $1-\varepsilon$. The exploration/exploitation phases are thus clearly separated. A typical parameter value might be $\varepsilon = 0.1$, but, actually, any fixed value of ε that is independent of T gives an expected cumulative regret that scales

Name of the policy	The score $\mu_{i,t}$
ε-GREEDY	$\overline{\mu}_{i,t-1}\mathbb{I}\{U_t \geq \varepsilon\}$
UCB	$1 \wedge \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{0.5\delta(t)}{N_{i,t-1}}}\right)$
UCB-V	$1 \wedge \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{2\overline{\sigma}^2_{i,t-1}\delta(t)}{N_{i,t-1}} + \frac{3\delta(t)}{N_{i,t-1}}}\right)$
UCB-KL	The largest $x \in [\overline{\mu}_{i,t-1}, 1]$ such that $N_{i,t-1}\mathrm{kl}\big(\overline{\mu}_{i,t-1}, x\big) \leq \delta(t)$
THOMPSON SAMPLING	An independent sample from $\mathrm{Beta}(\alpha, N_{i,t-1} - \alpha)$, where $\alpha = N_{i,t-1}\widetilde{\mu}_{i,t-1}$
MOSS	$1 \wedge \left(\overline{\mu}_{i,t-1} + \sqrt{\frac{0 \vee \log(T/(nN_{i,t-1}))}{N_{i,t-1}}}\right)$

TABLE 2.1: Some examples of popular index score for rewards in $[0,1]$. $\delta(t)$ is an exploration function, that can be of the form $\delta(t) = \zeta\log(t)$, with $\zeta > 1$, or $\delta(t) = \log(t) + 3\log\log(t)$, for $t \geq 3$, with $\delta(1) = \delta(2) = \delta(3)$. For TS, $\widetilde{\mu}_{i,t-1}$ can be $\overline{\mu}_{i,t-1}$ in practice. In theory, when rewards are not binary, it is maintained replacing each received reward $X_{i,t'}$ by a sample $Y_{i,t'} \sim \mathrm{Bernoulli}\big(X_{i,t'}\big)$.

linearly with T. Thus, since T is unknown, a value of ε decreasing as the experiment progresses is more judicious (Cesa-Bianchi and Fischer, 1998; Auer, Cesa-Bianchi, and Fischer, 2002), leading to very exploratory behavior at the beginning and very exploitative behavior at the end. We can also comment on why $\varepsilon = 0$ is a policy that will also have a linear regret in T. Let's consider to illustrate this fact the example of the Bernoulli bandit problem, with means in $(0,1)$. With a positive probability, after the initialization phase, the empirical average of the best arm will be 0 and that of all the others will be 1. From then on, the agent will never choose the best arm again, because there will always be another sub-optimal arm with a positive empirical average.

Optimistic policies

UCB, UCB-V and UCB-KL belongs to the family of optimistic policies, which proceed by maintaining a confidence region for each arm expected payoff. For an arm i, we construct the score $\mu_{i,t}$ as the supremum of the corresponding region. Thus defined, the arms score can also be interpreted as the largest statistically plausible mean value of the arm, given the currently available observation. Key ingredients in the construction of these algorithms (and also in their analysis) are concentration inequalities (see subsection 2.5.1), used to upper bound the estimation error with high probability (which is equivalent to constructing the confidence region). The terminology "optimistic" comes from the fact that for each arm i, the score $\mu_{i,t}$ is an upper confidence bound (UCB) on the true mean μ_i^* at round t, i.e., it is a slight overestimate of μ_i^*, with high probability. This overestimation is actually intended to counterbalance the fact that the empirical mean is negatively biased as an estimator for the true mean (Nie et al., 2017). This principle thus suggests to follow what seems to be the best arm, based on the optimistically constructed arm-scores. In other words, the

agent chooses the arm i_t by considering the most favorable environment among those that seem likely, i.e., chooses in an optimistic way.

Upper Confidence Bound (ucb) Introduced by Auer, Cesa-Bianchi, and Fischer (2002), the UCB algorithm relies essentially on the Hoeffding's inequality (see Theorem 5) to derive an upper bound on the regret (see Theorem 10). It derives for each arm a confidence interval around the empirical mean, and chooses at each round the arm with the highest UCB. Hoeffding's inequality allows a close form for the UCB of each arm i. More precisely, if we do not take into account that the final score is capped at 1 (which is a simple consequence of the assumption that the variables are in $[0,1]$), it is the sum of two terms: the empirical mean $\overline{\mu}_{i,t-1}$ and the exploration bonus $\sqrt{0.5\delta(t)/N_{i,t-1}}$, that is a high probability upper bound on the estimation error of $\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|$. The empirical mean is intended to guide the exploitation, measuring which arm is the best given the observed history. The exploration bonus is intended to guide exploration by indicating which arm has not been sampled much in the past. Indeed, the less an arm i has been sampled, the lower its counter $N_{i,t-1}$ will be and the bigger the bonus will be.

We will see later that the UCB algorithm can be analyzed relatively easily, and that its theoretical performance is very close to the lower bound given in Theorem 1. This makes UCB an ideal candidate for designing policies in many generalizations of MABs, such as stochastic combinatorial semi-bandits.

Upper Confidence Bound with Variance estimates (ucb-v) Auer, Cesa-Bianchi, and Fischer (2002) were the first to have the idea of exploiting the empirical variance of the arm in order to design the exploration bonus. Informally, as we have seen, an arm with a large variance should be explored more often than an arm with a small variance. Thus, the UCB policy that we presented in the previous paragraph explores too much, in the sense that the exploration bonus can be reduced in order to improve policy performance. The formal study of variance estimation in UCB type policies has been encouraged by the empirical superiority of such an approach (Audibert, Munos, and Szepesvári, 2009b). The resulting policy, known as the Upper Confidence Bound with Variance estimates (UCB-V), is based on the Bernstein's concentration inequality, sharper than that of Hoeffding (see Theorem 6). Since an approach considering a variance-dependent exploration bonus is not possible to construct for the agent (because the true variances are unknown), confidence regions around the empirical variances are also considered, giving in sum an approach based on an empirical Bernstein inequality, where the exploration bonus depends on the empirical variance (see Theorem 8). The gain in the exploration bonus compared to the UCB policy is immediately reflected on the regret bound (see Theorem 10).

Upper Confidence Bound using the Kullback–Leibler divergence (ucb-kl) The UCB-KL policy (also called KL-UCB in the literature (Garivier and Cappé, 2011)) uses an approach other than variance estimation in order to improve UCB, and is based on a KL confidence region for bounded random variables in $[0,1]$ (see Corollary 2). This policy and its analysis simultaneously appeared in Garivier and Cappé (2011) and Maillard, Munos, and Stoltz (2011), and was latter unified by Cappé et al. (2013). It is inspired by the seminal papers of Lai and Robbins (1985) and Burnetas and Katehakis (1996). As Maillard, Munos, and Stoltz (2011) indicate, it aims to achieve optimality guarantees in the sense of Burnetas and Katehakis (1996), i.e., to match

the asymptotic regret lower bound by deriving an upper bound of the form

$$R_T(\pi) \leq \sum_{i\in[n],\ \Delta_i>0} \frac{\log(T)(1+o(1))\Delta_i}{\mathrm{kl}(\mu_i^*, \mu_{i^*}^*)}.$$

To give an intuition about the algorithm, UCB-KL treats the problem as if it were a Bernoulli bandit one, and uses a confidence region specially adapted to this case. Noticing that binary rewards is a kind of worst case, the method extends to any variable in $[0,1]$. More concretely, as we will see in subsection 2.5.1, the crucial step for deriving a concentration inequality relies on a bound on the Moment-generating function (MGF) $\lambda \mapsto \mathbb{E}\left[e^{\lambda X}\right]$. This bound is of the same type as the one given for sub-Gaussian random variables (Definition 7). Thus, from random variables in $[0,1]$, we can go back to binary variables thanks to the following Lemma.

Lemma 1. *If X is a random variable in $[0,1]$ and $Y \sim \mathrm{Bernoulli}(\mathbb{E}[X])$, then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \mathbb{E}\left[e^{\lambda Y}\right].$$

Proof. If $\lambda \geq 0$, we have from $X \in [0,1]$,

$$\mathbb{E}\left[e^{\lambda X}\right] = \sum_{k\geq 0} \frac{\lambda^k \mathbb{E}\left[X^k\right]}{k!} \leq 1 + \sum_{k\geq 1} \frac{\lambda^k \mathbb{E}[X]}{k!} = 1 + \mathbb{E}[X]\left(e^{\lambda} - 1\right) = \mathbb{E}\left[e^{\lambda Y}\right].$$

In the case $\lambda < 0$, we first write $\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{-\lambda(1-X)}\right]e^{\lambda}$. Using the above result with $1 - X \in [0,1]$ and $-\lambda > 0$, we have $\mathbb{E}\left[e^{-\lambda(1-X)}\right] \leq \mathbb{E}\left[e^{-\lambda(1-Y)}\right]$. Thus, $\mathbb{E}\left[e^{-\lambda(1-X)}\right]e^{\lambda} \leq \mathbb{E}\left[e^{-\lambda(1-Y)}\right]e^{\lambda} = \mathbb{E}\left[e^{\lambda Y}\right]$. □

We can easily compare UCB-KL to UCB, since we can go from the first to the second by considering the quadratic divergence $2(p-q)^2$ instead of the Kullback-Leibler divergence $\mathrm{kl}(p,q)$. From Pinsker's inequality (giving $\mathrm{kl}(p,q) \geq 2(p-q)^2$ in this context), the KL confidence region is tighter, so the policy explores less, resulting in a better bound on the regret (see Theorem 10, one can also see Cappé et al. (2013) for an improved regret upper bound). Comparing UCB-KL to UCB-V is not straightforward, because UCB-KL treats the problem as if rewards were binary, leading to an overestimation of the variance. For instance, if rewards are deterministic in $(0,1)$, then UCB-V is better. On the other hand, if rewards are binary, then UCB-KL is better.

Thompson sampling

THOMPSON SAMPLING (TS) is an example of Bayesian randomized policy, introduced by Thompson (1933) (see also Thompson (1935)), simultaneously with the MAB problem. They have since been widely studied. Chapelle and Li (2011) provided a thorough empirical evaluation of the policy, highlighting its advantages. Agrawal and Goyal (2012b) gave an important first theoretical result, establishing a logarithmic regret bound. Latter, Kaufmann, Korda, and Munos (2012) proved that TS is actually matching the Lai and Robbins lower bound.

TS policy follows the Bayesian inference framework. The unknown distributions are parameterized with an assumed prior distribution. TS uses the prior distribution in each round with two phases: first it uses the prior distribution to sample a parameter, which is used to determine the action to play in the current round; second it uses the

feedback obtained in the current step to update the prior distribution to posterior distribution according to the Bayes' rule.

The use of Beta prior distribution is convenient as it is a conjugate distribution of the Bernoulli distribution. This allows easy computations of the posterior distribution after the observation of a Bernoulli realization. As for UCB-KL, we can extend the algorithm for non-binary rewards. The trick here is to update the model using another binary sample having the same mean as the true reward sample received. Indeed, in terms of expected regret, suffering the binary sample instead of the true reward is identical. Note that TS defines a family of algorithms as the choice of the priors is left to the user.

Policy for the distribution-free case

Let us mention here that there are also policies being optimal in the minimax sense, i.e., uniformly on all problems (they are matching the lower-bound from Theorem 3). The most well-known such policy is MOSS (Minimax Optimal Strategy in the Stochastic case) (Audibert and Bubeck, 2010; Degenne and Perchet, 2016a). This policy is very similar to UCB, involving a modified exploration function for each arm.

2.4.3 Low/high probability events

In designing the above policies (particularly those that are optimistic), we have referred to events that have a high probability of occurring. We give here a more precise meaning of this. For a round t, saying that an event $\mathfrak{A}_t$ holds with low probability means that $\mathbb{P}[\mathfrak{A}_t] = o(1/t)$ (and high-probability events are those whose complement holds with a low probability). We justify this now: By taking out events occurring with low probability, we wish to restrict the scope of possibilities to a high probability event in order to better guide the choice i_t. Specifically, the choice i_t will be relevant only under the high probability event $\neg\mathfrak{A}_t$, and will be meaningless otherwise. So if we decompose the regret as follows

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta_{i_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\Delta_{i_t}\mathbb{I}\{\mathfrak{A}_t\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\Delta_{i_t}\mathbb{I}\{\neg\mathfrak{A}_t\}\right],$$

then the first term can be bounded by $\max_{i\in[n]}\Delta_i\sum_{t=1}^{T}\mathbb{P}[\mathfrak{A}_t]$. We thus want

$$\sum_{t=1}^{T}\mathbb{P}[\mathfrak{A}_t] = o(\log(T)),$$

in order to be certain that it will be negligible compared to the second term, which we have seen must grow at least logarithmically (Theorem 1). One way to ensure this is to require $\mathbb{P}[\mathfrak{A}_t] = o(1/t)$.

2.5 Regret upper bound analysis

There are two types of upper bounds we can provide, exactly as for lower bounds. The first one is referred to as *gap-dependent*, because it depends explicitly on the underlying gaps Δ_i, $i \in [n]$. The second type is independent of the gaps, and is referred to as *gap-independent*, or *gap-free*. A gap-free bound gives more general guarantees as it would apply to several problem instances. However, the bound is then the worst case among these instances. On the other hand, a gap-dependent

bound depends on the problem instance and, in particular, helps to understand how fast the policy can learn in the easier problem instances (i.e., those with large gaps).

We will only give an analysis of the regret of the optimistic policies we presented above, for brevity's sake. As already mentioned, this kind of analysis is simpler to implement in the generalizations of the MAB setting to semi-bandit feedback and is therefore of primary interest for our presentation. We will also focus on the gap-dependent bounds. Indeed, we have just seen that gap-free bounds were not informative about the difficulty of the problem at hand, so we have chosen not to give priority to this kind of bound. Of course, when we have the opportunity to do so, we will indicate the extent to which policies behave from a minimax point of view. We refer the reader to the references given above for each policy for a case-by-case analysis.

Before going into more details about regret upper bounds, we first give an overview of the essential results allowing these bounds, namely concentration inequalities.

2.5.1 A tour of concentration inequalities

In this subsection, we prove the main results on the concentration of the empirical mean towards the true mean. Concentration is a phenomenon concerning the behavior of the distribution of the empirical mean. In order to quantify how the empirical mean converges to the true mean, we look at a region around the empirical mean that contains the true mean with high probability. The smaller this region is, the more the empirical mean has concentrated towards the true mean. In the MAB context, these regions are confidence intervals. They will not only serve to prove regret upper bounds, but are also useful in understanding how policies are designed. Indeed, for an arm i, the policy will choose the score $\mu_{i,t}$ as the UCB of the confidence interval around $\overline{\mu}_{i,t-1}$. Thus, the challenge is to build a confidence interval — containing the true mean with high probability — which is as tight as possible, because it would determine how the policy will explore. A region that is too big leads to a policy exploring too much. If the region is too small, the probability that the true mean does not belong to the region becomes non-negligible, and gives rise to poor performance as well.

In the whole subsection, we fix $i \in [n]$, $t \in \mathbb{N}^*$, and assume that $N_{i,t-1} \geq 1$. We start with two definitions. The first defines the cumulant-generating function of the random variable X_i as the logarithm of the MGF. The second defines the Legendre-Fenchel conjugate of this function. We will use the latter, evaluated as the empirical mean, to quantify the deviation from the true mean.

Definition 8 (Cumulant-generating function)**.** *The cumulant-generating function of X_i is*

$$\phi_i(\lambda) \triangleq \log \mathbb{E}\left[e^{\lambda X_i}\right],$$

defined for all $\lambda \in \mathbb{R}$ where it is finite, which includes at least $\lambda = 0$. If it exists on (λ_1, λ_2) containing 0, then ϕ_i is infinitely differentiable and is strictly convex on this interval (if X_i is not deterministic).

Definition 9 (Convex conjugate of ϕ_i)**.** *If ϕ_i is defined on an open interval (λ_1, λ_2) containing 0, then $\phi_i^* : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is the convex conjugate of ϕ_i, defined by*

$$\phi_i^*(z) \triangleq \sup_{\lambda \in (\lambda_1, \lambda_2)} (\lambda z - \phi_i(\lambda)) \geq 0.$$

It is strictly convex on $\{\phi^* < +\infty\}$*, and* $\phi_i(\mu_i^*) = 0$ *(by Jensen's inequality). In particular, it is continuous and decreasing (resp. increasing) on the interval* $(-\infty, \mu_i^*) \cap \{\phi_i^* < \infty\}$ *(resp.* $(\mu_i^*, +\infty) \cap \{\phi_i^* < \infty\}$*).*

Before moving on to the central result of this subsection, it is useful to notice a link between the sign of $z - \mu_i^*$ and and the side where the maximizer λ should be taken.

Lemma 2. *If* $z < \mu_i^*$*, then*

$$\phi_i^*(z) = \sup_{\lambda \in (\lambda_1, 0),\ \lambda z - \phi_i(\lambda) > 0} (\lambda z - \phi_i(\lambda)),$$

and if $z > \mu_i^*$*, then*

$$\phi_i^*(z) = \sup_{\lambda \in (0, \lambda_2),\ \lambda z - \phi_i(\lambda) > 0} (\lambda z - \phi_i(\lambda)).$$

Proof. We only prove the first part, since the second uses symmetric arguments. The function to maximize $\psi_z : \lambda \mapsto \lambda z - \phi_i(\lambda)$ is strictly concave. If there is a maximizer $\lambda \in (\lambda_1, \lambda_2)$, then the first order condition $\psi_z'(\lambda) = 0$ gives $\phi_i'(\lambda) = z < \mu_i^* = \phi_i'(0)$. Since ϕ_i' is increasing, this means that $\lambda < 0$. By strict concavity, we necessarily have $\psi_z(\lambda) > 0$, proving the lemma in this case. If no maximizer exists, since $\psi_z'(0) = z - \mu_i^* < 0$, we have by continuity that $\psi_z' < 0$ on (λ_1, λ_2), so ψ_z is decreasing and $\phi_i^*(z) = \lim_{\lambda \to \lambda_1} (\lambda z - \phi_i(\lambda))$. Since $\psi_z(0) = 0$, all the values $\lambda \in (\lambda_1, 0)$ are such that $\psi_z(\lambda) > 0$, while values $\lambda \in [0, \lambda_2)$ are such that $\psi_z(\lambda) \leq 0$, so the lemma is proved in this case as well. □

We now state the following theorem, which gives a bound on the probability that the gap between the empirical mean and the true mean is greater than a certain quantity. As we announced, the gap is measured with the function ϕ_i^*. The preceding Lemma 2 allows us to distinguish two cases, depending on whether the empirical mean is smaller or larger than the true mean. As usual, we can group these two cases together and end up with a probability multiplied by 2. The proof uses several ingredients, including a peeling to deal with the random counters $N_{i,t-1}$.

Theorem 4. *Assume that* ϕ_i *is finite on an open interval* (λ_1, λ_2) *containing* 0*. Then, for* $\delta > 1$*, the event*

$$\mathfrak{A}_t \triangleq \left\{ \mu_i^* > \overline{\mu}_{i,t-1} \text{ and } N_{i,t-1} \phi_i^*\left(\overline{\mu}_{i,t-1}\right) \geq \delta \right\}$$

holds with probability lower than $\lceil \delta \log(t) \rceil e^{1-\delta}$*. In a same way,*

$$\mathfrak{A}_t' \triangleq \left\{ \mu_i^* < \overline{\mu}_{i,t-1} \text{ and } N_{i,t-1} \phi_i^*\left(\overline{\mu}_{i,t-1}\right) \geq \delta \right\}$$

holds with probability lower than $\lceil \delta \log(t) \rceil e^{1-\delta}$*.*

Proof. As for Lemma 2, we only prove the first bound. We use a peeling argument in order to deal with the random counter $N_{i,t-1}$. For this, let $\gamma = \delta/(\delta - 1) > 1$, and consider the partition given by

$$\mathfrak{B}_{d,t} \triangleq \left\{ N_{i,t-1} \in (\gamma^{d-1}, \gamma^d] \right\},$$

for $d \in \left\{1, \ldots, \left\lceil \log_\gamma(t) \right\rceil \right\}$. We now want to bound the probability of the event $\mathfrak{A}_t \cap \mathfrak{B}_{d,t}$ and then use an union bound on $d \in \left\{1, \ldots, \left\lceil \log_\gamma(t) \right\rceil \right\}$. Since ϕ_i^* is decreasing

and continuous on the interval $(-\infty, \mu_i^*) \cap \{\phi_i^* < \infty\}$, we can define $x_i(N)$ for $N \geq 1$ as $x_i(N) = -\infty$ if $\phi_i^* < \delta/N$ on this interval, and as the unique solution $x \in (-\infty, \mu_i^*)$ to the equation $N\phi_i^*(x) = \delta$ otherwise. Consider now that $\mathfrak{A}_t \cap \mathfrak{B}_{d,t}$ holds. We have from

$$\gamma^d \phi_i^*\Big(\overline{\mu}_{i,t-1}\Big) \geq \gamma^d \delta / N_{i,t-1} \geq \delta,$$

that $\overline{\mu}_{i,t-1} \leq x\Big(\gamma^d\Big)$. Taking some $\lambda \in (\lambda_1, 0)$ such that $\lambda x\Big(\gamma^d\Big) > \phi_i(\lambda)$ thus gives

$$\exp\Big(\lambda N_{i,t-1}\overline{\mu}_{i,t-1} - \lambda N_{i,t-1} x\Big(\gamma^d\Big)\Big) \geq 1,$$

i.e.,

$$\begin{aligned}\exp\Big(N_{i,t-1}\Big(\lambda\overline{\mu}_{i,t-1} - \phi_i(\lambda)\Big)\Big) &\geq \exp\Big(N_{i,t-1}\Big(\lambda x\Big(\gamma^d\Big) - \phi_i(\lambda)\Big)\Big) \\ &\geq \exp\Big(\gamma^{d-1}\Big(\lambda x\Big(\gamma^d\Big) - \phi_i(\lambda)\Big)\Big).\end{aligned}$$

We can thus bound the probability of the event $\mathfrak{A}_t \cap \mathfrak{B}_{d,t}$ by the probability that the above inequality holds. The definition of ϕ_i implies that

$$\begin{aligned}&\mathbb{E}\Big[\exp\Big(N_{i,t-1}\Big(\lambda\overline{\mu}_{i,t-1} - \phi_i(\lambda)\Big)\Big)\Big] \\ &= \mathbb{E}\left[\prod_{u=1}^{t-1} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda)))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\prod_{u=1}^{t-1} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda)))\middle|\mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\prod_{u=1}^{t-2} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda)))\underbrace{\mathbb{E}[\exp(\mathbb{I}\{i_{t-1} = i\}(\lambda X_{i,t-1} - \phi_i(\lambda)))|\mathcal{F}_{t-1}]}_{=1}\right] \\ &= \cdots = 1.\end{aligned}$$

where we used that $\mathbb{I}\{i_t = i\} \in \mathcal{F}_t$, and $X_{i,t}$ is independent of $\mathcal{F}_t$. So from Markov inequality,

$$\begin{aligned}\mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{d,t}] \leq \mathbb{P}\Big[\exp\Big(N_{i,t-1}\Big(\lambda\overline{\mu}_{i,t-1} - \phi_i(\lambda)\Big)\Big) \geq \exp\Big(\gamma^{d-1}\Big(\lambda x\Big(\gamma^d\Big) - \phi_i(\lambda)\Big)\Big)\Big] \\ \leq \exp\Big(-\gamma^{d-1}\Big(\lambda x\Big(\gamma^d\Big) - \phi_i(\lambda)\Big)\Big).\end{aligned}$$

Notice that we can take the supremum of $\lambda x\Big(\gamma^d\Big) - \phi_i(\lambda)$ over $\lambda \in (\lambda_1, 0)$ such that $\lambda x\Big(\gamma^d\Big) > \phi_i(\lambda)$, thus giving using Lemma 2 with $x\Big(\gamma^d\Big) < \mu_i^*$ that

$$\mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{d,t}] \leq \exp\Big(-\gamma^{d-1}\phi_i^*\Big(x\Big(\gamma^d\Big)\Big)\Big) = e^{-\delta/\gamma}.$$

Putting everything together with an union bound, we have using $\log(\delta/(\delta-1)) \geq 1/\delta$, that

$$\mathbb{P}[\mathfrak{A}_t] \leq \sum_{d=1}^{\lceil \log_\gamma(t)\rceil} \mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{d,t}] \leq \Big\lceil \log_\gamma(t)\Big\rceil e^{-\delta/\gamma} = \lceil \delta \log(t)\rceil e^{1-\delta}.$$

□

As we've seen, the event $\mathfrak{A}_t$ defined in the previous theorem must be such that $\mathbb{P}[\mathfrak{A}_t] = o(1/t)$. This determines which choice for $\delta > 1$ we're going to take. Indeed, taking $\delta = \delta(t) = \log(t) + 3\log\log(t)$, for $t \geq 3$, with $\delta(1) = \delta(2) = \delta(3)$, gives for $t \geq 3$

$$\lceil \delta \log(t) \rceil e^{1-\delta} = e\Big[\log^2(t) + 3\log(t)\log\log(t)\Big] t^{-1} \log^{-3}(t) = o(1/t).$$

Notice also that $\delta = \delta(t) = \zeta \log(t)$ for $\zeta > 1$ also gives

$$\lceil \delta \log(t) \rceil e^{1-\delta} = e\Big[\zeta \log^2(t)\Big] t^{-\zeta} = o(1/t).$$

One last ingredient needed in order to effectively use $\mathfrak{A}_t$ in policy design is the actual construction of ϕ_i^*. This is where the hypothesis on the family of random variables under consideration has an impact. In fact, we are not going to use ϕ_i^* directly, but rather another function also measuring the gap to the true mean and which bounds inferiorly ϕ_i^*. Indeed, ϕ_i^* can be directly replaced by this lower bound in the event $\mathfrak{A}_t$, as this can't increase the probability of $\mathfrak{A}_t$. One way to build a lower bound on ϕ_i^* is to build an upper bound on ϕ_i, which coroborates the discussion before Lemma 1, where we mention the importance of upper bounding the MGF in order to have a concentration inequality. The simpler example of such upper bound is the one given in Definition 7, leading to the following well-known theorem, the Hoeffding's inequality.

Theorem 5 (Hoeffding's inequality, Hoeffding (1963))**.** *If X_i is κ^2-sub-Gaussian, then*

$$\mathbb{P}\left[\mu_i^* - \overline{\mu}_{i,t-1} \geq \sqrt{\frac{2\kappa^2\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t) \rceil e^{1-\delta(t)},$$

$$\mathbb{P}\left[\overline{\mu}_{i,t-1} - \mu_i^* \geq \sqrt{\frac{2\kappa^2\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t) \rceil e^{1-\delta(t)},$$

where $\delta(t) > 1$.

Proof. We prove the first inequality, the second is obtained symmetrically. The event can be rewritten as

$$\mathfrak{A}_t \triangleq \left\{\mu_i^* > \overline{\mu}_{i,t-1},\ 0.5N_{i,t-1}\Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big)^2/\kappa^2 \geq \delta(t)\right\}.$$

Thus, from Theorem 4, it is sufficient to prove that ϕ_i^* is lower bounded by $z \mapsto 0.5(\mu_i^* - z)^2/\kappa^2$. For this, we use the upper bound $\psi(\lambda) = \lambda\mu_i^* + \kappa^2\lambda^2/2$ on $\phi_i(\lambda)$, valid for all $\lambda \in \mathbb{R}$, thanks to the sub-Gaussianity assumption. Since we have $\psi^*(z) \leq \phi_i^*(z)$, it is sufficient to prove that $\psi^*(z) = 0.5(\mu_i^* - z)^2/\kappa^2$. This is a direct consequence of $\psi^*(z) = \sup_{\lambda\in\mathbb{R}}(\lambda z - \lambda\mu_i^* - \kappa^2\lambda^2/2)$, where the maximizer is $\lambda = (z - \mu_i^*)/\kappa^2$. □

Remark 5. *Note that our definition of the κ^2-sub-Gaussianity gives a bound on the MGF of the centered version of X_i. However, Theorem 5 also holds when for all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\lambda^2\kappa^2/2}$, with the same arguments.*

Using Hoefding's lemma (Proposition 2), we can state the following corollary.

Corollary 1. *If $X_i \in [0,1]$, then*

$$\mathbb{P}\left[\mu_i^* - \overline{\mu}_{i,t-1} \geq \sqrt{\frac{0.5\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t) \rceil e^{1-\delta(t)},$$

$$\mathbb{P}\left[\overline{\mu}_{i,t-1} - \mu_i^* \geq \sqrt{\frac{0.5\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t)\rceil e^{1-\delta(t)},$$

where $\delta(t) > 1$.

Due to its generality, Hoeffding's inequality has been widely used in MAB scenarios (the archetype is UCB). As mentioned before, a drawback of the bound is that it does not scale with the variance of X_i. If the variance is known, Bernstein's inequality can be used instead, which can yield significant improvements when the variance is small (as we will see in Chapter 4). In a more concrete way, in the case of bounded random variables in $[0,1]$, there exist better bounds on the MGF than the one given by sub-Gaussianity, and thus better concentration inequalities. Indeed, we provide the following lemma, giving that bounded random variables satisfies an MGF bound called *Bernstein's condition*, which will be useful in the derivation of the Bernstein's inequality.

Lemma 3 (Bernstein's condition for bounded random variables, see Vershynin (2009))**.** *If X is a random variable such that $|X - \mathbb{E}[X]| \leq c$, for some $c > 0$, then for all $\lambda \in (-3/c, 3/c)$,*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\mathbb{V}[X]/2}{1-|\lambda|c/3}}.$$

Proof. Let $|\lambda| < 3/c$, then

$$\begin{aligned}
\log\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] &= \log\left(1 + \sum_{k\geq 2}\frac{\lambda^k}{k!}\mathbb{E}\left[(X-\mathbb{E}[X])^k\right]\right) \\
&\leq \sum_{k\geq 2}\frac{\lambda^k}{k!}\mathbb{E}\left[(X-\mathbb{E}[X])^k\right] && \log(x) \leq x-1\ \forall x>0, \\
&\leq \lambda^2\mathbb{V}[X]\sum_{k\geq 2}\frac{|c\lambda|^{k-2}}{k!} && |X-\mathbb{E}[X]| \leq c \\
&\leq 0.5\lambda^2\mathbb{V}[X]\sum_{k\geq 2}\left(\frac{c|\lambda|}{3}\right)^{k-2} && k!/2 \geq 3^{k-2},\ \forall k \geq 2 \\
&= \frac{\lambda^2\mathbb{V}[X]/2}{1-|\lambda|c/3}.
\end{aligned}$$

□

We see that making the second order moment appear helps us to bound the MGF more precisely. We can also give a bound depending on the non-centered second order moment.

Lemma 4 (Non-centered one-sided Bernstein's condition)**.** *If X is a random variable such that $X \leq c$ a.s., for some $c \geq 0$, then for all $\lambda \in (0, 3/c)$,*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\mathbb{E}\left[X^2\right]/2}{1-\lambda c/3}}.$$

Proof. Notice that $g(x) = x^{-2}(e^x - 1 - x)$ is increasing on $\mathbb{R}$. We have $e^{\lambda X} = 1 + \lambda X + \lambda^2 X^2 g(\lambda X) \leq 1 + \lambda X + \lambda^2 X^2 g(\lambda c)$. Then, taking the expectation and using $\log(x) \leq x - 1\ \forall x > 0$, we have

$$\log\mathbb{E}\left[e^{\lambda X}\right] \leq \lambda\mathbb{E}[X] + \lambda^2\mathbb{E}\left[X^2\right]g(\lambda c).$$

Now, in the same way as previously, using $k!/2 \geq 3^{k-2}$, $\forall k \geq 2$, we have $g(\lambda c) = \sum_{k\geq 2}(\lambda c)^{k-2}/k! \leq 1/(2(1-\lambda c/3))$, giving the result. □

Remark 6. *Notice that we can have $c = 0$ in Lemma 4, in which case the result hold for any positive λ. This will be particularly useful in the proof of empirical Bernstein bound, using it with variables $-(X_i - \mu_i^*)^2 \leq c = 0$.*

We are now ready to state Bernstein's inequality. We also give a non-centered one-sided version. In a same way as for the Hoeffding's inequality, we prove the following Theorem 6 using the MGF bound given in Lemma 3.

Theorem 6 (Bernstein's inequality, Bernstein (1924))**.** *If $|X_i - \mu_i^*| \leq c$, then*

$$\mathbb{P}\left[\overline{\mu}_{i,t-1} - \mu_i^* \geq \frac{c\delta(t)}{3N_{i,t-1}} + \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t)\rceil e^{1-\delta(t)},$$

$$\mathbb{P}\left[\mu_i^* - \overline{\mu}_{i,t-1} \geq \frac{c\delta(t)}{3N_{i,t-1}} + \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t)\rceil e^{1-\delta(t)},$$

where $\delta(t) > 1$.

Proof. As previously, we only prove the first inequality (in fact, the second inequality can be found from the first replacing X_i by $-X_i$, and noticing that the upper bound given in the previous Lemma 3 remains valid, as $|-X_i - \mathbb{E}[-X_i]| \leq c$). We assume that $\mathbb{V}[X_i] > 0$, since otherwise $\overline{\mu}_{i,t-1} = \mu_i^*$ a.s., and the result trivially holds. We can derive the following lower bound on the function $\phi_i^*(z)$, $z > \mu_i^*$,

$$\psi^*(z) = \sup_{\lambda\in(0,3/c)}\left((z-\mu_i^*)\lambda - \frac{\lambda^2\mathbb{V}[X_i]/2}{1-|\lambda|c/3}\right) = \mathbb{V}[X_i]\sup_{\lambda\in(0,3/c)}\left(3u\lambda/c - \frac{\lambda^2/2}{1-\lambda c/3}\right),$$

where $u = c(z-\mu_i^*)/(3\mathbb{V}[X_i])$. The objective is concave in $(0, 3/c)$, so it is sufficient to find a critical point in $(0, 3/c)$. The critical points are given by

$$3u/c - \frac{\lambda}{1-\lambda c/3} - \frac{\lambda^2/6}{(1-\lambda c/3)^2} = 0, \text{ i.e., } \lambda^2 - \frac{6\lambda}{c} + \frac{18u}{c^2(1+2u)} = 0.$$

Solving the two degree polynomial gives $\lambda = 3c^{-1}\left(1 \pm \sqrt{1/(1+2u)}\right)$, so there is a single critical point in $(0, 3/c)$, given by $\lambda = 3c^{-1}\left(1 - \sqrt{1/(1+2u)}\right)$. Injecting this λ into the objective gives

$$\psi^*(z) = 9c^{-2}\mathbb{V}[X_i]h\left(\frac{c(z-\mu_i^*)}{3\mathbb{V}[X_i]}\right),$$

where $h(x) \triangleq 1 + x - \sqrt{1+2x}$ for $x \geq 0$. Notice that $h : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function, and $h^{-1}(x) = x + \sqrt{2x}$, $x \geq 0$. Applying Theorem 4, we get that

$$\mathfrak{A}_t \triangleq \left\{\mu_i^* < \overline{\mu}_{i,t-1} \text{ and } N_{i,t-1}\psi^*\left(\overline{\mu}_{i,t-1}\right) \geq \delta(t)\right\}$$

holds with probability lower than $\lceil \delta(t)\log(t)\rceil e^{1-\delta(t)}$. We can rewrite the condition $N_{i,t-1}\psi^*\left(\overline{\mu}_{i,t-1}\right) \geq \delta(t)$, using h^{-1}, as

$$\frac{c(\overline{\mu}_{i,t-1} - \mu_i^*)}{3\mathbb{V}[X_i]} \geq h^{-1}\left(\frac{c^2\delta(t)}{9\mathbb{V}[X_i]N_{i,t-1}}\right),$$

i.e.,

$$\overline{\mu}_{i,t-1} - \mu_i^* \geq \frac{c\delta(t)}{3N_{i,t-1}} + \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}} > 0.$$

□

Theorem 7 (Non-centered one-sided Bernstein's inequality)**.** *If $X_i \leq c \in [0, \infty)$, then*

$$\mathbb{P}\left[\overline{\mu}_{i,t-1} - \mu_i^* \geq \frac{c\delta(t)}{3N_{i,t-1}} + \sqrt{\frac{2\mathbb{E}[X_i^2]\delta(t)}{N_{i,t-1}}}\right] \leq \lceil \delta(t)\log(t)\rceil e^{1-\delta(t)},$$

where $\delta(t) > 1$.

Proof. We assume that $\mathbb{E}[X_i^2] > 0$, since otherwise $\overline{\mu}_{i,t-1} = \mu_i^*$ a.s., and the result trivially holds. We can derive the same lower bound as previously on the function $\phi_i^*(z)$, $z > \mu_i^*$, except that the variance is replaced by the second order moment $\mathbb{E}[X_i^2]$ using Lemma 4.

$$\psi^*(z) = \sup_{\lambda \in (0,3/c)} \left((z - \mu_i^*)\lambda - \frac{\lambda^2 \mathbb{E}[X_i^2]/2}{1 - \lambda c/3} \right) = \mathbb{E}\left[X_i^2\right] \sup_{\lambda \in (0,3/c)} \left(3u\lambda/c - \frac{\lambda^2/2}{1 - \lambda c/3} \right),$$

where $u = c(z - \mu_i^*)/(3\mathbb{E}[X_i^2])$. All the remaining of the proof is the same as in Theorem 6, giving the desired result. □

Since a priori knowledge of each arm variance is rarely available, this approach is not practical. A more suitable approach for MAB scenarios is to apply the previous Bernstein's inequality also to the empirical variance $\overline{\sigma}_{i,t-1}^2$. This results in a bound, which we will call the empirical Bernstein bound (Audibert, Munos, and Szepesvári, 2009b) that we now state.

Theorem 8 (Empirical Bernstein inequality)**.** *Assume $X_i \in [0, 1]$ and let $\delta(t) > 1$. With probability $1 - 3\lceil \delta(t)\log(t)\rceil e^{1-\delta(t)}$, we have*

$$\left|\mu_i^* - \overline{\mu}_{i,t-1}\right| \leq \frac{3\delta(t)}{N_{i,t-1}} + \sqrt{\frac{2\overline{\sigma}_{i,t-1}^2 \delta(t)}{N_{i,t-1}}}.$$

Proof. We do not directly apply Bernstein's inequality to the empirical variance, because the latter looks at the deviations from the empirical mean, whereas we wish to have the deviations from the true mean, in order to be able to apply the previous result replacing X_i by $-(X_i - \mu_i^*)^2$. Indeed, this is possible since $-(X_i - \mu_i^*)^2 \leq c = 0$. We thus get the two following inequalities, with probability $1 - 3\lceil \delta(t)\log(t)\rceil e^{1-\delta(t)}$,

$$\left|\mu_i^* - \overline{\mu}_{i,t-1}\right| \leq \frac{\delta(t)}{3N_{i,t-1}} + \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}} \tag{2.10}$$

$$-\widetilde{\sigma}_{i,t-1}^2 - (-\mathbb{V}[X_i]) \leq \sqrt{\frac{2\mathbb{E}[(X_i - \mu_i^*)^4]\delta(t)}{N_{i,t-1}}} \leq \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}}, \tag{2.11}$$

where $\widetilde{\sigma}_{i,t-1}^2 = (1/N_{i,t-1}) \sum_{t'=1}^{t} \mathbb{I}\{i_t = i\}(X_{i,t'} - \mu_i^*)^2$. We can notice that

$$\widetilde{\sigma}_{i,t-1}^2 - \overline{\sigma}_{i,t-1}^2 = \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^2,$$

so, using (2.11),

$$\begin{aligned}\mathbb{V}[X_i] - \overline{\sigma}^2_{i,t-1} &= \mathbb{V}[X_i] - \widetilde{\sigma}^2_{i,t-1} + \Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big)^2 \\ &\leq \sqrt{\frac{2\mathbb{V}[X_i]\delta(t)}{N_{i,t-1}}} + \Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big)^2. \end{aligned} \tag{2.12}$$

Letting $\ell = \delta(t)/N_{i,t-1}$, we now prove that

$$\sqrt{\mathbb{V}[X_i]} \leq \sqrt{\overline{\sigma}^2_{i,t-1}} + 1.8\sqrt{\ell}. \tag{2.13}$$

Since variables X_i are in $[0,1]$, we have $\sqrt{\mathbb{V}[X_i]} \leq 1/2$. If $\ell > 1/(3.6)^2$, then $1.8\sqrt{\ell} \geq 1/2 \geq \sqrt{\mathbb{V}[X_i]}$ and the inequality (2.13) holds. If $\ell \leq 1/(3.6)^2$, then, plugging (2.10) into (2.12) gives

$$\begin{aligned}\mathbb{V}[X_i] - \overline{\sigma}^2_{i,t-1} - \sqrt{2\mathbb{V}[X_i]\ell} &\leq \left(\frac{\ell}{3} + \sqrt{2\mathbb{V}[X_i]\ell}\right)^2 \\ &= \frac{\ell^2}{9} + \frac{2\ell\sqrt{2\mathbb{V}[X_i]\ell}}{3} + 2\mathbb{V}[X_i]\ell \\ &\leq \frac{\ell}{9\cdot(3.6)^2} + \frac{2\sqrt{2\mathbb{V}[X_i]\ell}}{3\cdot(3.6)^2} + \frac{\sqrt{\ell\mathbb{V}[X_i]}}{3.6} \\ &\leq \frac{\ell}{100} + 1.77\sqrt{\ell\mathbb{V}[X_i]}.\end{aligned}$$

This gives a second order polynomial inequality in $\sqrt{\mathbb{V}[X_i]}$, thus giving the bound $\sqrt{\mathbb{V}[X_i]} \leq 0.5\cdot 1.77\sqrt{\ell} + \sqrt{\overline{\sigma}^2_{i,t-1} + 0.8\ell} \leq \sqrt{\overline{\sigma}^2_{i,t-1}} + 1.8\sqrt{\ell}$, finishing the proof of (2.13). We finally plug (2.13) into (2.10) to get:

$$\Big|\mu_i^* - \overline{\mu}_{i,t-1}\Big| \leq \sqrt{2\overline{\sigma}^2_{i,t-1}\ell} + \Big(1.8\sqrt{2} + 1/3\Big)\ell < \sqrt{2\overline{\sigma}^2_{i,t-1}\ell} + 3\ell.$$

This proves the inequality of Theorem 8. □

Note that in the previous Theorem 8, we manage to quantify the error committed during the empirical estimation of the true mean, using an empirical quantity itself (the empirical variance). This is essential because it allows the agent to efficiently calculate and use the exploration bonus. However, once the choice i_t has been made using this empirical bonus, we wish to express a bound on the regret that can account for the advantage of using variance estimation. It may then be useful to further upper bound the empirical variance to bring out the true variance. This is the purpose of the next Proposition 3

Proposition 3 (Empirical variance concentration)**.** *Assume $X_i \in [0,1]$ and let $\delta(t) > 1$. With probability $1 - \lceil \delta(t)\log(t)\rceil e^{1-\delta(t)}$, we have*

$$\sqrt{\overline{\sigma}^2_{i,t-1}} \leq \sqrt{\mathbb{V}[X_i]} + \sqrt{\frac{0.5\delta(t)}{N_{i,t-1}}}.$$

Proof. We use the same notations than in the proof of Theorem 8 concerning $\widetilde{\sigma}^2_{i,t-1}$ and ℓ. Using Theorem 6 with X_i replaced by $(X_i - \mu_i^*)^2 \in [0,1]$, we can consider the

following event, which holds with probability $1 - \lceil \delta(t) \log(t) \rceil e^{1-\delta(t)}$.

$$\widetilde{\sigma}^2_{i,t-1} - \mathbb{V}[X_i] \leq \frac{\ell}{3} + \sqrt{2\mathbb{V}[X_i]\ell}.$$

The above inequality is obtained noticing that $\mathbb{V}\left[(X_i - \mu_i^*)^2\right] \leq \mathbb{V}[X_i]$. We assume that this event holds. Since $\widetilde{\sigma}^2_{i,t-1} - \overline{\sigma}^2_{i,t-1} = \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^2$, we can write

$$\overline{\sigma}^2_{i,t-1} - \mathbb{V}[X_i] \leq \overline{\sigma}^2_{i,t-1} - \mathbb{V}[X_i] + \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^2 \leq \frac{\ell}{3} + \sqrt{2\mathbb{V}[X_i]\ell}.$$

This is again a second order polynomial inequality in $\sqrt{\mathbb{V}[X_i]}$, giving

$$\sqrt{\mathbb{V}[X_i]} \geq \frac{-\sqrt{2\ell} + \sqrt{2\ell/3 + 4\sqrt{\overline{\sigma}^2_{i,t-1}}}}{2} \text{ or } \sqrt{\mathbb{V}[X_i]} \leq \frac{-\sqrt{2\ell} - \sqrt{2\ell/3 + 4\sqrt{\overline{\sigma}^2_{i,t-1}}}}{2}.$$

Since the second inequality is impossible, we have

$$\sqrt{\mathbb{V}[X_i]} \geq \frac{-\sqrt{2\ell} + \sqrt{2\ell/3 + 4\sqrt{\overline{\sigma}^2_{i,t-1}}}}{2} \geq \sqrt{\overline{\sigma}^2_{i,t-1}} - \frac{\sqrt{2\ell}}{2}.$$

□

The inequalities of Hoeffding and Bernstein are, as we have seen, obtained from a bound on the MGF. The tighter the bound, the more precise the inequality obtained. It is then legitimate to wonder if it is really necessary to bound this function, and if it is not possible to express it exactly. For some random variables (e.g., Gaussian random variables), the MGF is easy to express:

$$\text{if } X \sim \mathcal{N}(0, \mathbb{V}[X]),\ \mathbb{E}\left[e^{\lambda X}\right] = e^{\lambda^2 \mathbb{V}[X]/2}.$$

However, we often wish we could consider a family that is as general as possible. Actually, in the vast majority of MAB problems, we only know that the variables are bounded. Hoeffding's idea is then to look at what this boundedness can give as an upper bound on the MGF. He discovered (see Hoeffding's lemma, Proposition 2) that bounded variables MGF are bounded by that of a Gaussian, i.e., that they were sub-Gaussian. In Lemma 1, we have seen that the MGF of a random variable in $[0, 1]$ is bounded by that of a Bernoulli distribution. This seems better than Hoeffding's result because a Gaussian random variable is no longer bounded, whereas a Bernoulli one is. So the question is: can we simply express the MGF of a Bernoulli random variable? The answer is yes, and we'll even see that we can calculate ϕ_i^* easily in this case.

Proposition 4. *If X_i is a Bernoulli distribution, then $\phi_i^*(z) = \mathrm{kl}(z, \mu_i^*)$.*

Proof. We want to prove that for all $z \in \mathbb{R}$, $\sup_\lambda \lambda z - \log \mathbb{E}\left[e^{\lambda X_i}\right] = \mathrm{kl}(z, \mu_i^*)$. First order condition implies

$$z = \frac{\mathbb{E}\left[X_i e^{\lambda X_i}\right]}{\mathbb{E}[e^{\lambda X_i}]} = \frac{\mu_i^* e^\lambda}{\mu_i^* e^\lambda + (1 - \mu_i^*)}.$$

Thus,

$$\begin{aligned}\mathrm{kl}(z,\mu_i^*) &= z\log\left(\frac{z}{\mu_i^*}\right)+(1-z)\log\left(\frac{1-z}{1-\mu_i^*}\right)\\ &= z\log\left(\frac{e^\lambda}{\mu_i^*e^\lambda+(1-\mu_i^*)}\right)+(1-z)\log\left(\frac{1}{\mu_i^*e^\lambda+(1-\mu_i^*)}\right)\\ &= z\lambda-\log\Big(\mu_i^*e^\lambda+(1-\mu_i^*)\Big).\end{aligned}$$

□

We can thus adapt Theorem 4 together with the previous Proposition 4 and Lemma 1 to get the following Corollary 2.

Corollary 2. *If $X_i \in [0,1]$, then the event*

$$\mathfrak{A}_t \triangleq \Big\{\mu_i^* > \overline{\mu}_{i,t-1} \text{ and } N_{i,t-1}\mathrm{kl}\Big(\overline{\mu}_{i,t-1},\mu_i^*\Big) \geq \delta(t)\Big\}$$

holds with probability lower than $\lceil\delta(t)\log(t)\rceil e^{1-\delta(t)}$. In a same way,

$$\mathfrak{A}'_t \triangleq \Big\{\mu_i^* < \overline{\mu}_{i,t-1} \text{ and } N_{i,t-1}\mathrm{kl}\Big(\overline{\mu}_{i,t-1},\mu_i^*\Big) \geq \delta(t)\Big\}$$

holds with probability lower than $\lceil\delta(t)\log(t)\rceil e^{1-\delta(t)}$, with $\delta(t) > 1$.

2.5.2 Regret upper bounds

In this section, we focus on proving gap dependent regret bounds for the optimistic policies we previously introduced in section 2.4. To do so, we will make extensive use of the previously introduced concentration inequalities. The scheme of the proof is quite similar for the three algorithms, so we will unify the method used. More concretely, each algorithm considers, for each arm i a confidence region, which we will note $\mathcal{C}_{i,t}$. The score $\mu_{i,t}$ of the i arm is then the sup of the region $\mathcal{C}_{i,t}$. We recall in the Table 2.2 the definition of $C_{i,t}$ for each of the three algorithms. The tighter the confidence region around the empirical average will be, the tighter the bound will be. Since confidence intervals are only used on one side (because the policy is characterized by the sup of the n intervals), we need a second ingredient to control the behavior of the empirical mean on the left side of the true mean. This is the purpose of Theorem 9.

Theorem 9. *Let i be a sub-optimal arm. Let f be a non negative, decreasing and continuous function defined on the interval $[\mu_i^*,\mu_{i^*}^*]$, with $f(\mu_{i^*}^*)=0$. Then for all $t' \leq t$, $\varepsilon > 0$, the event*

$$\mathfrak{A}_{t,t'} \triangleq \Big\{N_{i,t-1}=t',\ f(\overline{\mu}_{i,t-1})\mathbb{I}\Big\{\overline{\mu}_{i,t-1}<\mu_{i^*}^*\Big\} < f(\mu_i^*)/(1+\varepsilon)\Big\}$$

holds with probability lower than $\exp(-t'r(\varepsilon,\mu_{i^*}^*,\mu_i^*))$, where $r(\varepsilon,\mu_{i^*}^*,\mu_i^*)$ is a positive constant dependent of $\varepsilon,\mu_{i^*}^*,\mu_i^*$. In particular, we have that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}\Big\{i_t=i,\ f(\overline{\mu}_{i,t-1})\mathbb{I}\Big\{\overline{\mu}_{i,t-1}<\mu_{i^*}^*\Big\} < f(\mu_i^*)/(1+\varepsilon)\Big\}\right]\\ &= \mathbb{E}\Big[\Big|\Big\{t\in[T],\ i_t=i,\ f(\overline{\mu}_{i,t-1})\mathbb{I}\Big\{\overline{\mu}_{i,t-1}<\mu_{i^*}^*\Big\} < f(\mu_i^*)/(1+\varepsilon)\Big\}\Big|\Big]\end{aligned}$$

Name of the policy	The region $\mathcal{C}_{i,t}$
UCB	$\left\{x \in [0,1],\ 2N_{i,t-1}\left(x - \overline{\mu}_{i,t-1}\right)^2 \leq \delta(t)\right\}$
UCB-V	$\left\{x \in [0,1],\ \left\lvert x - \overline{\mu}_{i,t-1}\right\rvert \leq \sqrt{\frac{2\overline{\sigma}^2_{i,t-1}\delta(t)}{N_{i,t-1}}} + \frac{3\delta(t)}{N_{i,t-1}}\right\}$
UCB-KL	$\left\{x \in [0,1],\ N_{i,t-1}\mathrm{kl}\left(\overline{\mu}_{i,t-1}, x\right) \leq \delta(t)\right\}$

TABLE 2.2: Confidence intervals for optimistic policies.

$$= \mathbb{E}[|\{t' \in [T],\ \exists t \geq t', \mathfrak{A}_{t,t'}\}|] \leq \sum_{t' \in [T]} \exp(-t' r(\varepsilon, \mu^*_{i^*}, \mu^*_i)) \leq 1/(1 - e^{-r(\varepsilon, \mu^*_{i^*}, \mu^*_i)})$$

is bounded by a constant dependent of $\varepsilon, \mu^*_{i^*}, \mu^*_i$.

Proof. We can define $x(\varepsilon)$ as the unique solution $x \in (\mu^*_i, \mu^*_{i^*})$ to the equation $f(x) = f(\mu^*_i)/(1+\varepsilon)$. Under the event $\mathfrak{A}_{t,t'}$, we necessarily have $\overline{\mu}_{i,t-1} > x(\varepsilon)$, and so $\exp\Big(N_{i,t-1}(\lambda\overline{\mu}_{i,t-1} - \phi_i(\lambda))\Big) \geq \exp(t'(\lambda x(\varepsilon) - \phi_i(\lambda)))$ for all $\lambda \geq 0$. By Markov inequality, the probability is thus bounded by $\exp(-t'(\lambda x(\varepsilon) - \phi_i(\lambda)))$, since the expectation of the LHS is 1:

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(N_{i,t-1}(\lambda\overline{\mu}_{i,t-1} - \phi_i(\lambda))\Big)\Big] \\
&= \mathbb{E}\left[\prod_{u=1}^{t-1} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda)))\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{u=1}^{t-1} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda)))\middle|\mathcal{F}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{u=1}^{t-2} \exp(\mathbb{I}\{i_u = i\}(\lambda X_{i,u} - \phi_i(\lambda))) \underbrace{\mathbb{E}[\exp(\mathbb{I}\{i_{t-1} = i\}(\lambda X_{i,t-1} - \phi_i(\lambda)))|\mathcal{F}_{t-1}]}_{=1}\right] \\
&= \cdots = 1.
\end{aligned}$$

Using the fact that $x(\varepsilon) > \mu^*_i$ in Lemma 2, we can take the supremum of $\lambda x(\varepsilon) - \phi_i(\lambda)$ over λ, which gives a probability bound of $\exp(-t'\phi^*_i(x(\varepsilon)))$. □

Theorem 10. *Assume that rewards are bounded in* $[0,1]$. *Let* $\delta(t) = \log(t) + 3\log\log(t)$, *for* $t \geq 3$, *with* $\delta(1) = \delta(2) = \delta(3)$, *and* $\varepsilon > 0$. *If* π *is the policy described by* UCB, *then*

$$R_T(\pi) \leq (1+\varepsilon) \sum_{i\ ,\Delta_i > 0} \frac{\log(T)(1+o(1))}{2\Delta_i}.$$

If π *is the policy described by* UCB-V, *then there exists a constant* c *such that*

$$R_T(\pi) \leq c \sum_{i\ ,\Delta_i > 0} \frac{\log(T)(\mathbb{V}[X_i] + \Delta_i + o(1))}{\Delta_i}.$$

If π *is the policy described by* UCB-KL, *then*

$$R_T(\pi) \leq (1+\varepsilon) \sum_{i\ ,\Delta_i > 0} \Delta_i \frac{\log(T)(1+o(1))}{\mathrm{kl}(\mu^*_i, \mu^*_{i^*})}.$$

Proof. Fixing any sub-optimal arm i, we can write

$$N_{i,T} = 1 + \sum_{t=n+1}^{T} \mathbb{I}\{i_t = i\}.$$

The goal is to bound the expectation of this sum. Indeed, this then induces a bound on the regret using the expression given in Remark 3. Let $t \geq n+1$ be some round where $i_t = i$. The confidence region of the optimal arm i^* satisfies

$$\mathbb{P}[\mu^*_{i^*} \in \mathcal{C}_{i^*,t}] \geq 1 - c' \lceil \delta(t) \log(t) \rceil e^{1-\delta(t)},$$

with c' being a universal constant, so we assume that this event holds at round t (the regret under its complementary is $o(\log(T))$). In addition, we can also assume that the event

$$\Big\{ f(\overline{\mu}_{i,t-1}) \mathbb{I}\Big\{ \overline{\mu}_{i,t-1} < \mu^*_{i^*} \Big\} \geq f(\mu^*_i)/(1+\varepsilon) \Big\} \tag{2.14}$$

holds at round t, since the regret under its complementary is bounded by a constant thanks to Theorem 9. We will chose f to be either $f(x) = \mathrm{kl}(x, \mu^*_{i^*})$ or $f(x) = (x - \mu^*_{i^*})^2$ depending if we consider the policy UCB-KL or UCB/UCB-V respectively. Using $\mu^*_{i^*} \in \mathcal{C}_{i^*,t}$ and the definition of $\boldsymbol{\mu}_t$, the following holds at round t

$$\mu_{i,t} = \sup_{j \in [n]} \sup \mathcal{C}_{j,t} \geq \sup \mathcal{C}_{i^*,t} \geq \mu^*_{i^*}.$$

Using that $\mu^*_{i^*} \geq \overline{\mu}_{i,t-1}$ (thanks to the event (2.14)), this rewrite as

$$\mathrm{kl}\Big(\overline{\mu}_{i,t-1}, \mu_{i,t}\Big) \geq \mathrm{kl}\Big(\overline{\mu}_{i,t-1}, \mu^*_{i^*}\Big), \quad \text{for UCB-KL}$$

and

$$\Big(\overline{\mu}_{i,t-1} - \mu_{i,t}\Big)^2 \geq \Big(\overline{\mu}_{i,t-1} - \mu^*_{i^*}\Big)^2, \quad \text{for UCB/UCB-V.}$$

The event (2.14) also gives a lower bound on the RHS of the two inequalities above:

$$\mathrm{kl}\Big(\overline{\mu}_{i,t-1}, \mu_{i,t}\Big) \geq \mathrm{kl}(\mu^*_i, \mu^*_{i^*})/(1+\varepsilon), \quad \text{for UCB-KL}$$

and

$$\Big(\overline{\mu}_{i,t-1} - \mu_{i,t}\Big)^2 \geq \Delta_i^2/(1+\varepsilon), \quad \text{for UCB/UCB-V.}$$

If $\Lambda_{f,\varepsilon,i}$ denotes the square root of the RHS in the above inequalities, we can use the definition of the confidence intervals to further upper bound $\Lambda_{f,\varepsilon,i}$ by some quantity $D_{i,t}$. When the policy is UCB-V, we first use the high probability bound on the empirical variance given by Proposition 3. We thus obtain the bound with $D_{i,t} = \sqrt{2\mathbb{V}[X_i]\delta(t)/N_{i,t-1}} + 4\delta(t)/N_{i,t-1}$. To sum up, it is sufficient to bound

$$\sum_{t=n+1}^{T} \mathbb{I}\{i_t = i,\ D_{i,t} \geq \Lambda_{f,\varepsilon,i}\}.$$

If the case of UCB-V, we use that $\mathbb{I}\{i_t = i,\ D_{i,t} \geq \Lambda_{f,\varepsilon,i}\}$ is upper bounded by

$$\mathbb{I}\Big\{ i_t = i,\ \sqrt{2\mathbb{V}[X_i]\delta(t)/N_{i,t-1}} \geq \Lambda_{f,\varepsilon,i}/2 \Big\} + \mathbb{I}\{i_t = i,\ 4\delta(t)/N_{i,t-1} \geq \Lambda_{f,\varepsilon,i}/2\},$$

and treat each case separately. So, let's note that all that's left to do is to bound a term of the form

$$\sum_{t=n+1}^{T} \mathbb{I}\Big\{i_t = i,\ a\delta(t)/\Lambda_{f,\varepsilon,i}^{\alpha} \geq N_{i,t-1}\Big\},$$

with $a \geq 0$ and $\alpha \in \{1, 2\}$. Noticing that $i_t = i$ implies that $N_{i,t} = N_{i,t-1} + 1$, by bounding $\delta(t)$ by $\delta(T)$, we get that the term above is lower than the number of integers in $[1, a\delta(T)/\Lambda_{f,\varepsilon,i}^{\alpha}]$, so is bounded by $a\delta(T)/\Lambda_{f,\varepsilon,i}^{\alpha}$. □

We can notice that the asymptotic bounds above give a logarithmic rate of regret when T is large. One question is what happens when T is relatively small compared to the terms $1/\Delta_i$. The minimax performance measure is then in this case more desirable. Results as Theorem 9 are not useful in this case, because although they allow to bound the regret under the corresponding event by a constant, the latter depends on the gaps Δ_i, and we rather wish to have a universal constant, even if it means having a greater multiplicative factor in front of the regret. We illustrate here how a logarithmic but not asymptotic bound on UCB can be converted into a minimax upper bound. This kind of conversion can also be derived for UCB-V and UCB-KL.

Theorem 11 (Minimax upper bound for UCB)**.** *Assume that rewards are bounded in* $[0, 1]$*. Let* $\delta(t) = \zeta \log(t)$*, with* $\zeta > 1$*. If* π *is the policy described by* UCB*, then*

$$R_T(\pi) \leq c + c' \sum_{i\ ,\Delta_i > 0} \frac{\log(T)}{\Delta_i},$$

where c, c' *are two universal constants.*
Letting $\eta > 0$*, we can thus write*

$$R_T(\pi) = \mathbb{E}\left[\sum_{t\in[T]} \mathbb{I}\{\Delta_{i_t} \leq \eta\}\Delta_{i_t}\right] + \mathbb{E}\left[\sum_{t\in[T]} \mathbb{I}\{\Delta_{i_t} > \eta\}\Delta_{i_t}\right] \leq T\eta + c + c'(n-1)\frac{\log(T)}{\eta}.$$

Taking $\eta = (c'(n-1)\log(T))^{1/2} T^{-1/2}$ *gives*

$$R_T(\pi) \leq c + 2(c'(n-1)\log(T)T)^{1/2}.$$

Proof. The proof is similar to the one given for Theorem 10. For a round t, we can write

$$\mathbb{P}[\forall i \in [n],\ \mu_i^* \in \mathcal{C}_{i,t}] \geq 1 - 2n\lceil\delta(t)\log(t)\rceil e^{1-\delta(t)}.$$

We now assume that this high probability event holds at round t (the regret under the complementary is bounded by a constant independent of Δ_i), and use it to have $\mu_{i,t} \geq \mu_{i^*}^*$, i.e., $\sqrt{2\delta(t)/N_{i,t-1}} = \operatorname{diam} \mathcal{C}_{i,t} \geq \mu_{i,t} - \mu_i^* \geq \Delta_i$, since both $\mu_{i,t}$ and μ_i^* belongs to $\mathcal{C}_{i,t}$. Exactly as for Theorem 10, we obtain the main term rate. □

2.6 Some extensions

This section presents two specific extensions of the MAB problem found in the literature, namely, *stochastic linear bandits* and *budgeted multi-armed bandit*. We chose to focus on these two extensions because they find applications in stochastic combinatorial semi-bandits, as we will see in upcoming chapters.

2.6.1 Stochastic linear bandits

An interesting variant of MAB is the stochastic linear bandit problem, introduced by Auer (2002). In this setting, the set of arms is no longer $[n]$ but a subset $\mathcal{X}$ of $\mathbb{R}^n$. The set $\mathcal{X}$ is fixed and revealed to the agent. Pulling an arm $\mathbf{x} \in \mathcal{X}$, the agent observes a noisy reward whose expected value is the dot product between $\mathbf{x}$ and an unknown parameter $\boldsymbol{\theta}^* \in \mathbb{R}^n$,

$$X_{\mathbf{x}_t,t} = \mathbf{x}_t^\intercal \boldsymbol{\theta}^* + \eta_t,$$

where $\mathbf{x}_t$ is the arm pulled at round t and η_t is a centered noise affecting the reward observation at round step t. Usually, more details about the noise models are specified in each problem definition.

The above setting is intended to impose a linear structure on the reward received by the agent. It becomes particularly interesting in applications where the number of arms is very important compared to n. Indeed, we can see that the rewards are linearly parameterized in a fixed n-dimensional space that does not depend on the number of arms, contrary to classical MABs. On the other hand, in linear bandits, pulling one arm gives information on the parameter $\boldsymbol{\theta}^*$ and thus indirectly, on the reward value of the other arms. Therefore, the estimate of the average reward in MAB is replaced here by the estimate of the components of $\boldsymbol{\theta}^*$. To put it plainly, by modelling such a structure, the agent will be able to take advantage not only of the dimension reduction imposed by the latent parameter, but also of the assumption of linearity of the reward.

In this context, to guide the sample allocation strategy, the agent has to rely on the estimation of $\boldsymbol{\theta}^*$ obtained from the observed rewards. More precisely, after observing a sequence of rewards $X_{\mathbf{x}_1,1}, \dots, X_{\mathbf{x}_t,t}$, an ordinary least squares (OLS) estimator can be defined for the sample parameter $\boldsymbol{\theta}^*$ as

$$\hat{\boldsymbol{\theta}}_t \triangleq V_t^{-1} \left(\sum_{t' \in [t]} \mathbf{x}_{t'} X_{\mathbf{x}_{t'},t'} \right),$$

where the matrix $V_t \triangleq \sum_{t' \in [t]} \mathbf{x}_{t'} \mathbf{x}_{t'}^\intercal$ is called the *design matrix*, and is an agglomeration of the pulling strategy up to round t. The OLS estimate enjoys a series of interesting properties, such as concentration inequalities, that we are not going to detail here, but that can be found for instance in Abbasi-Yadkori, Pál, and Szepesvári (2011). It is by using the concentration inequalities on $\hat{\boldsymbol{\theta}}_{t-1}$ that the bandit policies can be designed, and select which arm $\mathbf{x}_t$ to play. Specifically, we can construct a confidence ellipsoid whose center is the empirical estimate $\hat{\boldsymbol{\theta}}_{t-1}$ and which containing the true parameter $\boldsymbol{\theta}^*$ with high probability. The confidence ellipsoid is usually with respect to the euclidean norm associated to the design matrix V_{t-1}, i.e., of the form

$$\mathcal{C}_t \triangleq \left\{ \boldsymbol{\theta} \in \mathbb{R}^n, \ \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t-1} \right\|_{V_{t-1}}^2 \leq \delta(t) \right\}.$$

In fact, since the uncertainty here comes from the n components of $\boldsymbol{\theta}^*$, the width of the ellipsoid in a certain direction is determined by the precision of the estimates corresponding to that direction. Thus, the goal is to be able to shrink as fast as possible the volume of the confidence ellipsoid as the agent observes more and more samples. Similarly to the MAB setting, the policies for cumulative regret minimization can use a UCB strategy: they construct upper-bounds on the arms values using the margins of the confidence ellipsoids. Then, according to the OFU principle, the policies pull

the arms with the largest index. This is a direct adaptation of the UCB policy, that start from the empirical estimate on the arm value $\mathbf{x}^\intercal \hat{\boldsymbol{\theta}}_{t-1}$, to which an exploration bonus is added, based on the uncertainty of estimating the reward of $\mathbf{x}$. Specifically, here the uncertainty is captured in the width of the confidence ellipsoid in direction $\mathbf{x}$. In other word, the UCB of the arm $\mathbf{x}$ is

$$\mathbf{x}^\intercal \hat{\boldsymbol{\theta}}_{t-1} + \sqrt{\delta(t)} \|\mathbf{x}\|_{V_{t-1}^{-1}}.$$

This arm index construction has been introduced in Auer (2002). The good empirical performance of an index-based policy following this type of construction was shown in Li, Chu, et al. (2010).

The construction of the arm indices can also be seen as a bilinear optimization problem

$$\max_{\mathbf{x}\in\mathcal{X},\ \boldsymbol{\theta}\in\mathcal{C}_t} \mathbf{x}^\intercal \boldsymbol{\theta} = \max_{\mathbf{x}\in\mathcal{X}} \left(\mathbf{x}^\intercal \hat{\boldsymbol{\theta}}_{t-1} + \sqrt{\delta(t)} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right).$$

We can finish by noticing that when $\mathcal{X}$ is large, the above optimization problem is usually difficult to solve (it is in general NP-Hard, see Atamtürk and Gómez (2017)). Thus, methods of the type TS may be preferred (Agrawal and Goyal, 2012b; Abeille and Lazaric, 2017), because, in this context, the above optimization problem becomes a linear program over $\mathcal{X}$, since instead of maximizing over $\boldsymbol{\theta}$, we rather sample this vector from a prior belief, and then simply optimize $\max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\intercal \boldsymbol{\theta}$.

2.6.2 Budgeted Multi-armed Bandit

In budgeted MABs, to play an arm, the agent needs to pay a cost while receiving the reward. In this new context, the agent targets at maximizing the cumulative reward under a budget constraint for the total costs. This setting has been first studied by Tran-Thanh, Chapman, et al. (2012), Ding, Qiny, et al. (2013), Vanchinathan et al. (2015), and Xia, Ding, et al. (2016). The budgeted MAB problem can be formally defined as follows. At round t, the agent pulls an arm $i_t \in [n]$, receives a random reward $X_{i_t,t}$, and pays a random cost $C_{i_t,t}$. Usually, we assume that both $X_{i,t}$ and $C_{i,t}$ are supported in $[0,1]$ for all arm $i \in [n]$, and have mean μ_i^* and c_i^* respectively. The agent can keep pulling until the budget, B, runs out. B is a positive number and does not need to be known to the agent in advance. Note that we do not assume that the rewards of an arm are independent of its costs. The agent wants to follow a policy π that minimizes the regret, which is usually defined as the differences between F_B^*, the maximum expect cumulative reward that a pulling policy can obtain within a budget B when the reward/cost distributions of all the arms are known, and the expected reward that the policy π obtains under the budget constraint B. Formally,

$$R_B(\pi) \triangleq F_B^* - F_B(\pi),$$

where, for a policy π selecting arm i_t at round t,

$$F_B(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} X_{i_t,t} \right],$$

τ_B being the random round at which the remaining budget becomes negative: if $B_t \triangleq B - \sum_{t' \le t} C_{i_{t'},t'}$, then $B_{\tau_B - 1} \ge 0$ and $B_{\tau_B} < 0$.

It is hard to find the optimal policy for the above setting. Even for the *offline* setting in which the reward and cost of each arm are deterministic and known, the

problem is an unbounded knapsack problem, which is NP-hard (Lueker, 1975). We can nevertheless define a quasi-optimal policy, in the sense that its regret is bounded by a constant. The quasi-optimal policy can simply be described as an index policy: as long as the budget is not exhausted, the agent pulls the arm $i = \arg\max \mu_i^*/c_i^*$. As in standard MAB, this *offline* policy can inspire *online* policies. Indeed, in a learning context, the constant due to the non-optimality of the offline policy will be relatively small compared to the exploration/exploitation error that appears when B is sufficiently large (more precisely, this error is logarithmic in B). Thus, to be more specific, a UCB type policy can be used: the agent chooses in each round t the arm maximizing the ratio $\mu_{i,t}/c_{i,t}$, where $\mu_{i,t}$ is a UCB on the true mean μ_i^*, and $c_{i,t}$ is a lower confidence bound (LCB) on the true expected cost c_i^*. Note here that the OFU principle results in an overestimation of the gains, but in an underestimation of the costs.

Chapter 3

General Framework for Semi-Bandit Feedback

This chapter is an introduction to the framework of *stochastic combinatorial multi-armed bandits with semi-bandit feedback* (CMAB), called more concisely *stochastic (combinatorial) semi-bandits*. We organize it in the following way: first of all, we formulate the CMAB setting in a general way, thus allowing a large number of problem to be expressed. Then we give concrete examples of real-life use to motivate the framework introduced. Finally, we present some general results that will be useful to obtain regret upper bound for CMAB.

3.1 The stochastic combinatorial semi-bandits problem

The extension of MAB to the CMAB framework can be described in a few words: the agent can now choose more than one arm in the same round. We call *super-arm* or *action* a set of arm that the agent is allowed to play. The set of actions is denote by $\mathcal{A} \subset \mathcal{P}([n])$ and is called the *action space*. This makes it possible to model more complex situations, where the actions given to the agent are complex and can be broken down into several small actions (the arms). In addition to the trade-off between exploration and exploitation, CMAB must also deal with the exponential explosion of possible actions that makes the exploration of all actions unfeasible. The feedback received by the agent at round t is composed of the individual feedback of each arm in the chosen super-arm $A_t \in \mathcal{A}$, i.e., the agent observes $\mathbf{e}_{A_t} \odot \mathbf{X}_t = (X_{i,t}\mathbb{I}\{i \in A_t\})_{i\in[n]}$ (it observes only the outcomes of played arms in one round of play). The reward obtained is a function of the individual outcomes. In the majority of cases, the most natural reward to consider is *linear* with respect to the chosen action (Kveton, Wen, Ashkan, and Szepesvari, 2015a), i.e., it is the sum of the individual outcomes:

$$\mathbf{e}_{A_t}^\intercal \mathbf{X}_t = \sum_{i\in A_t} X_{i,t}.$$

Remark 7. *We can note that the distribution $\mathbb{P}_{\mathbf{X}}$ is no longer reduced to the collection of the marginals. Indeed, since several arms are played together in one round, the entire joint distribution of rewards plays a role. In particular, the correlations between the arms matter. We will come back to this point in Chapter 7.*

In the literature, the CMAB setting was first considered through some specific instances of the problem. A number of studies considered simultaneous plays of any subset of m arms among the n arms that are available (Anantharam, Varaiya, and Walrand, 1987; Caro and Gallien, 2007; Liu, Liu, and Zhao, 2011). More complex scenarios include:

- The matching bandit problem (Gai, Krishnamachari, and Jain, 2010), where the set of arms is the set of edges in a fixed bipartite graph, and $\mathcal{A}$ is the set of matchings in this bipartite graph (we recall that a matching is a set of edges without common vertices).
- The online shortest path problem (Liu and Zhao, 2012; Talebi, Zou, et al., 2013), where the set of arms is the set of directed edges in a fixed directed acyclic graph with a given source and destination, and $\mathcal{A}$ is the set of path from the source to the destination.

We will give in sections 3.2 more scenarios of application for the CMAB setting. In the following subsection we formalize more precisely the CMAB problem, incorporating a generalization that will allow us to consider more application cases: we will assume that the feedback obtained by the agent is random. This semi-bandit setting is known as *Semi-Bandits with Probabilistically Triggered Arms* (Chen, Wang, and Yuan, 2016; Wang and Chen, 2017), and is denoted CMAB-T.

3.1.1 Probabilistically triggered arms

We consider the setting in which actions may trigger arms probabilistically. In this context, we denote the action space $\mathcal{S}$, which is no longer necessarily a subset of $\mathcal{P}([n])$, but rather an external set that we allow to be infinite. At round t, the agent selects an action $S_t \in \mathcal{S}$, based on the feedback history from the previous rounds and a possible extra source of randomness. Then, the environment draws an independent sample $\mathbf{X}_t \sim \mathbb{P}_{\mathbf{X}}$, $\mathbf{X}_t \in \mathbb{R}^n$. Then, a random subset of arms $A_t \subset [n]$ are triggered. The set A_t depends on S_t and $\mathbf{X}_t$, but may have additional randomness. In other words, we assume that A_t is drawn independently from a distribution $D_{\text{trig}}(S_t, \mathbf{X}_t)$. The outcomes of each triggered arm is observed as the feedback to the agent, i.e., $\mathbf{e}_{A_t} \odot \mathbf{X}_t$ is observed. The agent finally obtains a reward $\rho(S_t, A_t, \mathbf{X}_t)$. Quantities $\mathcal{S}, D_{\text{trig}}, \rho$ are known to the agent.

This setting notably captures the CMAB framework (without probabilistically triggered arms) described at the beginning of the section, taking $\mathcal{S} = \mathcal{A}$ and D_{trig} the Dirac measure at S_t. More generally, it also encodes randomized policies, taking $\mathcal{S}$ equal to the set of probability measures on $\mathcal{A}$, and $D_{\text{trig}}(S_t, \mathbf{X}_t) = S_t$. It is sometimes convenient to consider randomization as "undergone" by the agent, meaning that it is encoded in A_t whereas S_t is deterministic. This is the case when the good properties of the policy (such as being a solution to a certain optimization problem) are carried by the distribution S_t used by the agent to generate A_t rather than by A_t itself. Conversely, it is sometimes more practical to consider randomization as "controlled" by the agent, meaning that it is encoded in the action S_t, as in the previous chapter. This is the case when the random action S_t bears the good properties of the policy (i.e., S_t satisfies some criterion, such as being solution to some optimization problem, but is generated with a distribution that is not easy to describe otherwise). In any case, we can extend the definition of the filtration $(\mathcal{F}_t)$ as

$$\mathcal{F}_t \triangleq \begin{cases} \sigma(U_1) & \text{if } t = 1 \\ \sigma(U_1, A_1, \mathbf{e}_{A_1} \odot \mathbf{X}_1, \dots, U_{t-1}, A_{t-1}, \mathbf{e}_{A_{t-1}} \odot \mathbf{X}_{t-1}, U_t) & \text{if } t \geq 2. \end{cases}$$

When we do not want to consider the randomization in the filtration, we use $\mathcal{H}_t$ instead, that we call *history*:

$$\mathcal{H}_t \triangleq \begin{cases} \sigma(\emptyset) & \text{if } t = 1 \\ \sigma(U_1, A_1, \mathbf{e}_{A_1} \odot \mathbf{X}_1, \dots, U_{t-1}, A_{t-1}, \mathbf{e}_{A_{t-1}} \odot \mathbf{X}_{t-1}) & \text{if } t \geq 2. \end{cases}$$

We will give in section 3.3 more examples of application for CMAB-T, where D_{trig} also depends on the sample $\mathbf{X}_t$.

Remark 8. *In the following, we will use the notation A_t to designate the set of arms that is sampled in round t, and S_t the action that is chosen by the agent. In the classical CMAB setting (without probabilistically triggered arms), we can notice that A_t and S_t coincides. In this case, we will often use the notation A_t only, for the sake of simplicity. Nevertheless, in some contexts it is convenient to distinguish the set A_t (the coordinates of the vector $\mathbf{X}_t$ which are revealed to the agent) from the action S_t (the action chosen by the agent). This is notably the case when the action S_t can be described more succinctly than by the set A_t of arms associated with it.*

3.1.2 The (approximation) regret

The performance of a policy is measured by its regret, which is the difference in expected cumulative reward between always playing the best action and playing actions selected by the policy. Note, however, that even when the agent can compute exactly the expected reward $\mathbb{E}[\rho(S, A, \mathbf{X})]$ for any $S \in \mathcal{S}$, i.e., even when the learning process of the distribution $\mathbb{P}_{\mathbf{X}}$ is completed, maximizing the quantity $\mathbb{E}[\rho(S, A, \mathbf{X})]$ for $S \in \mathcal{S}$ can be hopeless. In some cases, there exist efficient approximation algorithms, that we call oracles, outputting $S \in \mathcal{S}$ such that

$$\mathbb{E}[\rho(S, A, \mathbf{X})] \geq \alpha \sup_{S' \in \mathcal{S}} \mathbb{E}[\rho(S', A', \mathbf{X})],$$

where $\alpha \in (0, 1]$ is an approximation ratio, and $A \sim D_{\text{trig}}(S, \mathbf{X})$, $A' \sim D_{\text{trig}}(S', \mathbf{X})$. Under an α-approximation oracle, the benchmark cumulative reward should be the α fraction of the optimal reward. Thus to evaluate the performance of a learning policy π, we use the notion of approximation regret (Kakade, Kalai, and Ligett, 2009; Streeter and Golovin, 2009; Chen, Wang, and Yuan, 2016), defined as follows.

Definition 10 (Approximation regret)**.** *The T-round α-approximation regret of a learning policy π that selects action $S_t \in \mathcal{S}$ at round t is defined as*

$$R_{T,\alpha}(\pi) \triangleq \mathbb{E}\left[\sum_{t \in [T]} \Delta(S_t)\right],$$

where the approximation gap is defined as

$$\Delta(S) \triangleq 0 \vee \left(\alpha \sup_{S' \in \mathcal{S}} \mathbb{E}[\rho(S', A', \mathbf{X})] - \mathbb{E}[\rho(S, A, \mathbf{X})]\right).$$

We also use the shortcut $\Delta_t \triangleq \Delta(S_t)$. When $\alpha = 1$, meaning that one can efficiently maximize $\mathbb{E}[\rho(S, A, \mathbf{X})]$, we denote the regret as $R_T = R_{T,1}$.

Definition 11 (Reward function)**.** *Generally, the expected reward $\mathbb{E}[\rho(S, A, \mathbf{X})]$ is a function of some unknown parameter vector $\mathbf{w}^*$ related to the distribution $\mathbb{P}_{\mathbf{X}}$, in which case we use the notation $r(S; \mathbf{w}^*) \triangleq \mathbb{E}[\rho(S, A, \mathbf{X})]$. The agent thus aims to minimize the (approximation) regret by learning this unknown vector. For instance, the mean vector $\boldsymbol{\mu}^*$ can be this parameter vector when $\mathbb{P}_{\mathbf{X}} = \otimes_{i \in [n]} \text{Bernoulli}(\mu_i^*)$. The function r is called the reward function. A first example (that does not involves probabilistically triggered arms) is when the reward is linear: we have $r(A; \boldsymbol{\mu}) \triangleq \mathbf{e}_A^\intercal \boldsymbol{\mu}$, because $\mathbb{E}[\mathbf{e}_A^\intercal \mathbf{X}] = \mathbf{e}_A^\intercal \boldsymbol{\mu}^*$.*

3.1.3 Other types of feedback

There exists other type of feedback in combinatorial bandits (Audibert, Bubeck, and Lugosi, 2011), that include: full information, in which the player observes the outcomes of all arms, and bandit, in which the player only observes the final reward but no outcome of any individual arm. More complicated feedback dependences are also considered in Mannor and Shamir (2011). In the whole thesis, we focus on the semi-bandit feedback we introduced above.

3.2 Real world applications of CMAB

CMAB problems are driven by a wide range of real-world situations. Although many models are quite simplistic to fully solve real world scenarios, they provide a very good benchmark for providing effective solutions to these problems. In this section, we will describe some examples of applications of the CMAB framework. We recall that the feedback is semi-bandit: the agent observes the outcomes of the arms belonging to the chosen action $S_t = A_t$. The last two applications (about maximum coverage) are examples where it is convenient to have $S_t \neq A_t$, although they do not involves probabilistically triggered arms.

In each of the following settings (except for the last two), note that we have the choice to take a linear reward function depending on the need and the context. When we choose a linear reward function, we take $\alpha = 1$ in the definition of the regret, because there is an offline algorithm that can optimize exactly the reward function in a time polynomial in n. For example, the maximum weighted bipartite matching problem (also called the assignment problem) can be solved in polynomial time using the Hungarian algorithm (Kuhn, 1955). For the shortest path problem, many efficient algorithms exists, and the most important are the Dijkstra's algorithm (Dijkstra et al., 1959) (with non-negative edge weight) and the Bellman–Ford algorithm (see e.g., Jukna and Schnitger (2016)).

3.2.1 More on the matching bandit problem

There is considerable interest in the development of mechanisms to access the digital spectrum for more efficient use (as we saw earlier for the application of MAB to cognitive radios). Indeed, cognitive radio networks, characterised by higher levels of autonomy, intelligence and learning, are expected to play an important role in this area. The matching bandit setting described in the previous section can model a type of problem that can occur when optimizing the use of the communication network. More precisely, the bipartite graph might represent, on the one hand, the users, and on the other hand, the channels in which these users want to transmit. At each round, a central instance (the agent) must assign each user to a channel, so that there is no collision (i.e., if an user is assigned to some channel, there is no other conflicting user on that channel). Notice that we thus implicitly assume that there are more channels than users. Each edge of the graph is associated to a random outcome, and the goal for the agent is to choose its matching so as to maximize the sum over the edges belonging to the matching of the corresponding outcomes (notice, the reward function is thus linear here). These outcomes represent the quality of the transmission made by the user. The difference between the outcomes encodes the geographical dispersion of the users: the channels are more or less accessible depending on many factor related to the user behaviour. We give in Figure 3.1 an example of a bipartite graph encoding an instance of the setting just described.

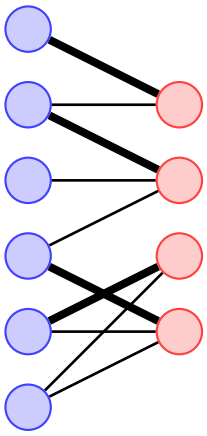

FIGURE 3.1: Example of a bipartite graph considered in the matching bandit problem. Red nodes are users, blue nodes are channels. Edges represents the base arms. When an edge between a user and a channel is present in the graph, it means that the user can transmit through that channel. An example of matching is formed by the bold edges.

Another problem falling into the same type of combinatorial optimization is the following: the agent is an online dating site, and it is earning at each round an uncertain gain for each match made in that round. In addition to the dating market, we can imagine other scenarios where finding a maximum weight match is a goal for the agent.

3.2.2 More on the online shortest path problem

Routing is the mechanism by which paths are selected in a network to route data from a sender to one or more recipients. Broadly, routing is performed in many types of networks, including circuit-switched networks, such as the public switched telephone network (PSTN), and computer networks, such as the Internet. Its performance is important in decentralized networks, i.e., where information is not distributed by a single source, but exchanged between independent nodes. Here, we consider, as introduced in the previous section, a multi-hop communication network represented by a directed graph. Assume that we have to send a stream of packets from the source node to the destination node. At each round, a packet is sent along a chosen route connecting the source to the destination. Depending on the network traffic (and other factors, such as signal strength in wireless networks), each edge in the network may have a different delay (the random outcome is usually minus the delay, to transform the minimization problem into a maximization one), and the total delay the packet suffers on the chosen route is the sum of delays of the edges composing the route (i.e., again, the reward function is linear).

The setting can also benefit when searching for the shortest route in a road network. With increasing urbanisation and the growth of cities with several million inhabitants, road networks are under great pressure. The problem of traffic allocation concerns the allocation of routes on a network that seeks to minimise a cost defined by the instance of the problem (normally travel time is used). In order to assign the shortest route to a user, the route planner (the agent) can use the outputs of the previous recommendations (such as the speed of the vehicle along the route) to optimize the proposed itinerary.

3.2.3 Dynamic assortment

We consider a price-taking retailer (the agent) that has n products to sale. We assume that each product has a fixed known marginal profit resulting from selling it. For instance, the marginal profit can be the price at which the product is sold minus the marginal cost paid by the agent for offering the product. At each round, a customer arrives, with some unknown random valuation vector over products. Then, the agent offers any subset of m products (due to display space constraints, the retailer can offer at most m products simultaneously), and the customer buys an offered product if and only if its valuation is greater than its price. The agent is interested in maximizing the total profit (revenue minus cost) from sales over T rounds, where T denotes the total number of customers that arrive during the selling season. This setting have been studied for instance in Sauré and Zeevi (2013).

3.2.4 Web page optimization

Building web pages that match user preferences is a major issue in a variety of situations: online advertising is an important application, as is the customization of home pages or search engine results (Lagrée, Vernade, and Cappe, 2016). Learning how to place items in multi-position displays or lists is a task that can be integrated into the semi-banded framework. To be more precise, our agent here is the designer of the web page in question. At the beginning of each round (a round can be for instance an hour, a day, or simply a user reaching the page), the agent chooses a design for the web page: it chooses which elements (clickable) and where to place them. At the end of the round, the agent observes the outcomes corresponding to the number of clicks (or the click-through-rate) on each element placed, and can update the design for a new round. The agent's goal is to maximize the total number of clicks. In this setting, arms are pairs (item, position), and super-arms are a set of m arms with pairwise disjoint positions and items. Depending on the setting considered, several types of reward functions can be used: we can for example associate to each pair (item, position) an outcome, in which case we have a linear reward function. Note that this setting is similar to the matching bandit setting, where the items form one part of the bipartite graph, and the positions form the second part.

3.2.5 Crowdsourcing

Crowdsourcing has received considerable attention in recent years and many applications have been successful, such as gathering information quickly during a disaster, performing tasks that are difficult to automate and need to be solved by human workers, conducting large-scale surveys or contributing to scientific projects. There are at least two types of crowdsourced tasks: micro-tasks, uniformly priced at a few cents, and expert tasks, where the employer (the agent) has much more control over the selection of individual workers, and must also take into account the potentially very heterogeneous and higher costs (workers often charge 10 to 50 per hour). To address the specific challenges of crowdsourcing experts, the agent can see workers as arms. This set of workers is usually determined by an open call for participation by the employer, to which qualified and available workers respond. Assigning a single task to a worker can be regarded as pulling the arm. This incurs a cost that is set by the worker, and the assignment produces an outcome. The agent has a total budget of B to spend on crowdsourcing the tasks, and wishes to maximize the overall sum of the outcomes (notice, we thus here also fall into the budgeted bandit setting described in the end of the previous chapter). Last but not least, the agent would also like to

have completed the tasks in a minimum amount of time. Thus, at each round, the agent pays an additional fixed cost, modeling the time. In this context, the agent should therefore assign tasks to several workers in a single round, in order to go faster, while taking into account the efficiency of each agent. A similar setting have been considered in Tran-Thanh, Stein, et al. (2014).

3.2.6 Real-time strategy games

We have already seen that MCTS approaches such as UCT are useful for exploring game trees. These techniques have been successful for games with a moderate branching factor. For real-time strategy (RTS) games, where the magnitude of the branching factor is combinatorial due to the fact that multiple units can be issued actions simultaneously, a CMAB approach may be preferable. Indeed, RTS games are adversarial games, generally simulating battles or complex interactions between a large number of units (workers, military units, ...), that pose a significant challenge to both human and artificial intelligence (because of the enormous action and state space, and because they are real time). In addition, some RTS games are also partially observable and non-deterministic, which further motivates the use of CMABs. More specifically, the use of MABs in MCTS alorithms was useful to explore and evaluate the quality of the actions available. As in the context of RTS games, the available actions are broken down into several local actions, the MAB approach naturally translates into a CMAB approach: the agent benefits from the estimation of a local action (i.e., an arm) in each super-arm containing it. The internal structure of the setting is thus exploited. The above setting have been considered in Ontanón (2013).

It should be noted, however, that the reward function may be difficult to capture in many RTS games, and that it is generally not linear, as specific combinations of local actions may lead to specific rewards. On the other hand, the assumption of semi-bandit feedback is also debatable. Indeed, the outcome of a played arm is not always observed individually, especially when other arms interfere with the perceived outcome (which can be, depending on the situation, a gain of resources, the destruction of strategic buildings, ...). Nevertheless, it can be argued that only the current state of the game (which does not change during MC simulations) influences the outcome that a certain arm produces.

3.2.7 Probabilistic maximum coverage bandit

The probabilistic maximum coverage (PMC) bandit problem is the bandit version of the maximum coverage problem (Hochbaum, 1997). As stated previously, it is an example where it is practical to distinguish the chosen action S_t from the set of arms A_t that is associated to that action. The problem can be described as follows: consider a bipartite graph $G = (L, R, E)$, with $E \subset L \times R$, where each edge ij is an arm and has an unknown probability $\mu^*_{ij} \in [0, 1]$ associated to it. The action space $\mathcal{S}$ is the collection of subsets of L of size m. The random vector $\mathbf{X} \in \mathbb{R}^E$ is sampled from $\otimes_{ij \in E} \text{Bernoulli}\left(\mu^*_{ij}\right)$. Each time an action S is chosen, the set of arms $A = E \cap (S \times R) = \{ij \in E : \ i \in S\}$ is observed. The reward is the cardinality of the set $\{j : \ ij \in A, \ X_{ij} = 1\}$, i.e., is the number of right hand vertex that are covered by a node from the action S (see Figure 3.2). Notice thus that the reward function (i.e., the expected reward) is

$$r(S; \boldsymbol{\mu}) \triangleq \sum_{j \in R} \left(1 - \prod_{i \in S: \ ij \in E} (1 - \mu_{ij}) \right).$$

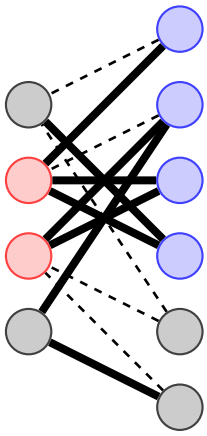

FIGURE 3.2: Example of a bipartite graph considered in the PMC bandit problem. Red nodes are those in the action S. Edges ij represented with bold lines are those with $X_{ij} = 1$, and the others, represented with dash lines, are such that $X_{ij} = 0$. A bold edge adjacent to a red node makes the other node of that edge blue. Thus, blue nodes are those covered by the red ones. The reward is the number of blue nodes.

As a function of S, the reward function r is monotone, meaning that $r(S; \boldsymbol{\mu}) \geq r(S'; \boldsymbol{\mu})$ for $S' \subset S$, and *submodular* (Fujishige, 2005), meaning that for any pair of sets S and S', we have

$$r(S; \boldsymbol{\mu}) + r(S'; \boldsymbol{\mu}) \geq r(S \cup S'; \boldsymbol{\mu}) + r(S \cap S'; \boldsymbol{\mu}).$$

Submodularity will be further studied in chapters 5 and 6. We can say here that the problem of maximizing r, for a certain parameter $\boldsymbol{\mu}$, over subsets of L of size m, is inapproximable within a factor better than $1 - 1/e$, unless P = NP (Feige, 1998). The GREEDY algorithm for maximum coverage chooses sets according to one rule: at each stage (and until the solution is of size m), add the L node to the current solution which covers the largest number of uncovered elements. It can be shown that this algorithm achieves an approximation ratio of $1 - 1/e$ (Hochbaum, 1996). In conclusion, in the above Definition 10 of the approximation regret, we take $\alpha = 1 - 1/e$ for the PMC bandit problem.

PMC bandit have been studied by Chen, Wang, and Yuan (2013), and more recently by Merlis and Mannor (2019). Several applications of PMC exist (McGregor and Vu, 2019), such as sensor allocation (Krause and Guestrin, 2007), information retrieval (Anagnostopoulos et al., 2015), ad placement (Radlinski, Kleinberg, and Joachims, 2008), influencer marketing (Kempe, Kleinberg, and Tardos, 2015) or blog monitoring (Saha and Getoor, 2009). We give here one real-world application of the PMC framework introduced above.

Content delivery network A content delivery network (CDN) (Pathan, Buyya, and Vakali, 2008) is a collaborative collection of network elements spanning the Internet and arranged for more effective delivery of the content. In a CDN, content is replicated over several mirrored Web servers in order to perform transparent and effective delivery of content to the end users. The mirrored servers are strategically placed at various locations to deal with sudden spikes in the requests. A critical aspect in the operation of a CDN is how contents are placed on surrogate servers.

Ideally, content is placed on a set of surrogate servers that should be chosen to maximize the total number of requests served. For each server hosting the content, the list of users who have made a request to access the content from that server is known, and may change on the fly. The same user can send a request to several servers when possible. The PMC bandit setting can be used to sequentially select the set of servers (L nodes). The rounds represent time steps, and in each round t, the agent selects a set of m servers (m stands for a constraint on the maximum number of servers that can be used). Then, for each selected server, the list of users (R nodes) submitting a request to that server within the current round is received (modeled by $X_{ij,t}$). The goal is to maximize the total number of users covered (summing over all rounds).

3.2.8 Weighted maximum coverage bandit

For some applications, in the PMC bandit problem introduced above, the sets of R nodes covered by a selected L node is not random. On the other hand, the gain associated with each node can be random (no longer equal to 1). This kind of scenario can happen for example in the above CDN example: let's imagine that surrogate servers are connected to several infrastructures (deterministically), but that the number of requests linked to these infrastructures is random, and does not depend on the type of server connected to the infrastructure. In this kind of context, the arms are no longer the edges, but the R nodes, and the outcomes are the random gain (i.e., the number of requests) associated to them. It is thus easier to learn how to optimize the coverage because the number of parameters to learn is no longer $|L| \times |R|$, but only $|R|$. Moreover, the set R can in principle be much smaller than before, because users are grouped into infrastructures (one can imagine that the total number of infrastructures can potentially be much smaller than the total number of users). If S is a set of m nodes from L, and if $\mathbf{w} \in \mathbb{R}_+^R$ is a parameter vector standing for the mean of the R node outcomes, then the reward function is defined as

$$r(S; \mathbf{w}) \triangleq \sum_{j \in R} w_j \mathbb{I}\{S \text{ is connected to } j\}.$$

This function is still monotone and submodular in S, and for the same reasons as for the PMC bandit, we take $\alpha = 1 - 1/e$ in Definition 10.

3.3 Real world applications of CMAB-T

In this section, we give some concrete examples of use of the CMAB-T setting. We recall that S_t and A_t are always different in this case: A_t is an element of $\mathcal{P}([n])$, while S_t belongs to an action space $\mathcal{S}$ that may not be a subset of $\mathcal{P}([n])$.

3.3.1 Cascading bandits

Cascading bandits have been studied in Kveton, Szepesvari, et al. (2015), Kveton, Wen, Ashkan, and Szepesvari (2015a), and Li, Wang, et al. (2016), and can be introduced as follows. There are n arms, associated with independent Bernoulli outcomes. An action is any ordered sequence of arms of cardinality m. Playing an action means that the outcomes of the arms are revealed one by one following the sequence order until certain stopping condition is satisfied. The feedback is the outcomes of revealed arms and the reward is a function of these arms. There are two main forms for this problem:

- The disjunctive form: the agent stops when the first 1 is revealed and gains reward of 1, or it reaches the end and gains reward 0. The reward function is thus of the form
$$r(S;\boldsymbol{\mu}) \triangleq 1 - \prod_{i\in S}(1-\mu_i).$$
The triggered set of arms A is a prefix set of the action $S = (s_1, \dots, s_m)$: $A = \{s_1, \dots, s_k\}$, such that k is the first in S with $X_{s_k} = 1$.

- The conjunctive form: the agent stops when the first 0 is revealed (and receives reward 0) or it reaches the end with all 1 outcomes (and receives reward 1). The reward function is thus of the form
$$r(S;\boldsymbol{\mu}) \triangleq \prod_{i\in S}\mu_i.$$
The triggered set of arms A is $\{s_1, \dots, s_k\}$, such that k is the first in S with $X_{s_k} = 0$.

Cascading bandits can be used to model online recommendation and advertising. In this context, the agent recommend a list of m items to a user (such as web pages). The user examines the recommended list from the first item to the last, and selects the first attractive item (i.e., click on it for web search), or don't select any item if none is attractive. This is an example of the disjunctive form with an outcome 1 for each attractive item, and a reward 1 if a click occurs. Another application is network routing reliability. In this context, the action S represents of a path to route data from a sender to a recipients. The goal for the agent is to be able to route the data using a path that does not contain a defective arm (i.e., a broken edge). This is an example of the conjunctive form with an outcome 0 for each broken arm, and a reward of 1 when the data was successfully transmitted (i.e., no broken edge were present in the selected path).

Let us note finally that the functions r described here are maximized exactly by taking S containing the m arms having the largest means (we can thus take $\alpha = 1$ in the regret).

3.3.2 Weighted probabilistic maximum coverage bandit

We can mix the two maximum coverage bandit problems introduced in the previous section. In this new setting, we have random weights associated with R nodes on the top of the PMC bandit setting. Even if both problems were instances of CMAB, their association becomes a CMAB-T problem, because the random weights associated with R nodes are only observed if there is a selected L node covering it: there is therefore randomness in the feedback the agent receives. More precisely, our set of arms here is E, but it should be noted that each arm $ij \in E$ has two outcomes: a Bernoulli variable $X_{ij} \in \{0, 1\}$ encoding if the node i covers the node j, and another variable $W_j \in \mathbb{R}_+$ encoding the weight associated with the node j. We assume that $\mathbb{P}_{(\mathbf{W},\mathbf{X})} = \mathbb{P}_{\mathbf{W}} \otimes \left(\otimes_{ij\in E}\mathbb{P}_{X_{ij}}\right)$, i.e., weights are assume to be independent from $\mathbf{X}$ (note, however, that the weights are not assumed to be mutually independent). When an action $S_t \subset L$ is played at round t, the feedback received is

$$(X_{ij,t}\mathbb{I}\{i \in S\}),\ (W_{j,t}\mathbb{I}\{\exists i \in S,\ X_{ij,t} = 1\}).$$

We can thus see that the feedback set associated with $\mathbf{X}_t$ is not random (it is the same as for the PMC bandit problem), whereas the feedback set associated with $\mathbf{W}_t$ is: the randomness is given by $\mathbf{X}_t$ itself.

Using the independence assumption we can see that the reward function is

$$r(S; \boldsymbol{\mu}, \mathbf{w}) \triangleq \sum_{j \in R} w_j \left(1 - \prod_{i \in S:\ ij \in E} (1 - \mu_{ij})\right).$$

As before, r is submudular, and we can define the approximation regret with $\alpha = 1 - 1/e$.

The types of scenarios in which this setting can occur are the same as those in the PMC bandit problem, but where not every node has the same importance to be covered. In the context of CDN, this can happen, for example, when users are heterogeneous (they do not have the same type of subscription, they are willing to pay more for access to content, ...).

3.3.3 Online influence maximization

Influence maximization (IM) is the problem of finding a small set of most influential nodes in a social network so that their aggregated influence in the network is maximized. Online influence maximization (OIM) is the bandit version of this problem (Chen, Wang, and Yuan, 2016; Wen, Kveton, Valko, et al., 2017). We refer the reader to Chapter 6 for further details about OIM. In IM (Kempe, Kleinberg, and Tardos, 2015), we are given a directed graph $G = (V, E)$, where V and E are sets of vertices and edges respectively. Each edge ij represents a connections between the two users i and j (e.g., i follows j in the social network). An underlying diffusion model D governs how information spreads in G. More precisely, D is a probability distribution on subgraphs[1] G' of G. Several types of distributions D exist in the literature, and the two best known are as follows:

- Independent cascade model (IC): the random subgraph $G' = G_{\mathbf{W}}$ is

  $$G_{\mathbf{W}} \triangleq (V, \{ij \in E :\ W_{ij} = 1\}),$$

  where $\mathbb{P}_{\mathbf{W}} \triangleq \otimes_{ij \in E} \text{Bernoulli}\left(w^*_{ij}\right)$.

- Linear threshold model (LT): the random subgraph G' is $G_{\mathbf{W}}$, with $\mathbb{P}_{\mathbf{W}} \triangleq \otimes_{j \in V} \text{Multinoulli}\left(\left(w^*_{ij}\right)_{i:\ ij \in E}\right)$, i.e., for every $j \in V$, select at most one of its incoming edges at random, such that edge ij is selected with probability w^*_{ij}, and no edge is selected with probability $1 - \sum_{i:\ ij \in E} w^*_{ij} \geq 0$.

Given some seed set S, the spread $\sigma(S; \mathbf{w})$ is defined as the expected number of S-reachable nodes in $G' \sim D$ (i.e., the number of nodes in G' that are reachable from some node in S), where D is a distribution parameterized by $\mathbf{w}$ as in the IC and LT models. IM aims to find S that is a solution to

$$\max_{|S|=m} \sigma(S; \mathbf{w}). \tag{3.1}$$

Although IM is NP-hard under standard diffusion models — i.e., IC and LT — σ is a monotone submodular function of S. In particular, Kempe, Kleinberg, and Tardos

[1] A subgraph of a graph G is obtained by removing some edges from G.

(2015) provide a greedy algorithm with an approximation ratio of $\alpha = 1 - 1/e - \varepsilon$ for all $\varepsilon > 0$.

For OIM, the distribution D is unknown and is parameterized by an unknown vector $\mathbf{w}^*$. D must be learned over time by repeated influence maximization tasks: at each round t, m seed nodes S_t are selected, the influence propagation of S_t is observed and the reward is the number of nodes activated in that round. The agent wants to repeat this process to accumulate as much reward as possible. More precisely, the set of arms is the set of edges E, the outcomes are $\mathbf{W}_t$ and the feedback set is $A_t = \left\{ij \in E : \ S_t \overset{\mathbf{W}_t}{\rightsquigarrow} i\right\}$, where $G_{\mathbf{W}_t} \sim D$, and where the predicate $S_t \overset{\mathbf{W}_t}{\rightsquigarrow} i$ holds if, in the graph $G_{\mathbf{W}_t}$, there is a forward path from a node in S_t to the node i. In other word, A_t is the set of edges reachable from S_t in $G' = G_{\mathbf{W}_t}$. The reward $\rho(S, A, \mathbf{W})$ is the number of nodes that is reached from S through G', and the expected reward is exactly the influence spread $\sigma(S; \mathbf{w}^*)$.

3.4 General technical results: toward proving regret upper bounds

In this section, now that we have reviewed many applications of the CMAB and CMAB-T settings, we will begin to see how an upper bound on the regret of a policy can be proven. In all the thesis, we are only interested in the rate that regret has up to a multiplicative universal factor. In other words, we are interested in proving upper bound of the form

$$R_T(\pi) = \mathcal{O}(f(T)),$$

with f being a function that may depend on the input of the problem. By $\mathcal{O}$, we mean that the ratio $R_T(\pi)/f(T)$ must be bounded by a universal constant for all $T \in \mathbb{N}^*$. It should be noted that this section contains some fairly technical derivations, and that the reader who wants to go through the thesis quickly can skip the rest of the chapter to go to the next Chapter 4. The results that are presented here will find applications throughout the thesis, and are stated in a general enough way so that they can be used in a large number of applications, that may go beyond the scope of the thesis.

In this section, for the sake of generality, and to cover also the budgeted bandit settings, we assume that the horizon T may be random. Let's recall the definition of the counters:

$$\forall i \in [n], \ \forall t \geq 1, \ N_{i,t-1} = \sum_{t'=1}^{t-1} \mathbb{I}\{i \in A_{t'}\}.$$

There are also some other definitions that will be useful for the analysis and the regret bound expression.

Definition 12 (Triggering probabilities)**.** *For an arm $i \in [n]$ and an action $S \in \mathcal{S}$, we define the probability that action S triggers arm i as*

$$p_i(S) \triangleq \mathbb{P}[i \in A],$$

where $A \sim D_{trig}(S, \mathbf{X})$. In particular, since $S_t \in \mathcal{F}_t$, we have

$$p_i(S_t) = \mathbb{P}[i \in A_t | \mathcal{F}_t].$$

Several other quantities can be defined from the gaps and $p_i(S)$:

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

$$\forall i \in [n],\ \Delta_{i,\max} \triangleq \max_{S \in \mathcal{S},\ p_i(S)>0,\ \Delta(S)>0} \Delta(S).$$

As a convention, if there is no action S such that $p_i(S) > 0$ and $\Delta(S) > 0$, we define $\Delta_{i,\min} = +\infty$ and $\Delta_{i,\max} = 0$. We also define $\Delta_{\min} \triangleq \min_{i \in [n]} \Delta_{i,\min}$ and $\Delta_{\max} \triangleq \max_{i \in [n]} \Delta_{i,\max}$. We define the maximum number of arms that can be triggered together with some arm i as

$$m_i \triangleq \max_{S \in \mathcal{S},\ p_i(S)>0} \sum_{j \in [n]} \mathbb{I}\{p_j(S) > 0\}.$$

Finally, we let $b_i(S_t) \geq 0$ be some quantity depending on i and S_t, which we can choose according to our needs.

3.4.1 How to prove regret bounds?

As we saw in the previous chapter, the regret analysis of an MAB policy is based on high probability events. It will be the same for CMAB-T policies. To recall, at a round t, the possibility space for this round is divided into several mutually exclusive and exhaustive events, and the regret is filtered against each of these events. We thus obtain a sum of several *event-filtered regrets*. The goal is to upper bound each of these terms. One term dominates the others when T is large. It is called the *leading term* and it summarizes the regret upper bound rate. In a sense, therefore, the other events aim to restrict the possibilities — without their filtered regret being too large — in order to gather the right conditions to prove the leading term upper bound. Generally, the leading term is obtained by bounding Δ_t by an error term that is easier to apprehend when summed over $t \in [T]$. The way to bound Δ_t by an error term depends on the policy that is under consideration. As for index-based MAB policies, there is a usual form for CMAB-T policies, that can be described as follows:

- Consider an estimate $\boldsymbol{\mu}_t$ of the true mean at round t.
- Plug $\boldsymbol{\mu}_t$ into an *oracle* that α-approximately maximizes $S \mapsto r(S; \boldsymbol{\mu}_t)$.

Under the right conditions, the estimate $\boldsymbol{\mu}_t$ is "close enough" to $\boldsymbol{\mu}^*$ so that we can bound Δ_t by

$$\alpha \sup_{S' \in \mathcal{S}} r(S'; \boldsymbol{\mu}_t) - r(S_t; \boldsymbol{\mu}^*).$$

Using the definition of the oracle, this is further bounded by

$$r(S_t; \boldsymbol{\mu}_t) - r(S_t; \boldsymbol{\mu}^*). \tag{3.2}$$

The type of error term that we obtain in the end therefore depends on how "close enough" $\boldsymbol{\mu}_t$ is to $\boldsymbol{\mu}^*$. More precisely, it is necessary to arbitrate between a $\boldsymbol{\mu}_t$ fairly optimistic so as to provide the first bound on Δ_t, and a $\boldsymbol{\mu}_t$ not too optimistic so as to reduce the gap (3.2) as much as possible. If, to simplify, we leave optimism aside, the choice of the empirical average $\boldsymbol{\mu}_t = \overline{\boldsymbol{\mu}}_{t-1}$ seems obvious. The gap (3.2) is thus controlled by the concentration of $\overline{\boldsymbol{\mu}}_{t-1}$ to $\boldsymbol{\mu}^*$. Therefore, we generally refer to the error term as a *bonus*, even if it is not explicitly used in the policy (or if the policy is

not even optimistic). The purpose of this section is to see what kind of regret bound we can get depending on the type of bonus we have.

An example of bonus building There are many ways to proceed in order to build a bonus. Generally speaking, there are two basic ingredients at work. First, a smoothness relation to link the variation of the objective function $r(S;\boldsymbol{\mu})$ to the variation of the parameter $\boldsymbol{\mu}$. This smoothness relation provides a bound on (3.2). Then, a concentration inequality, in order to limit the variation of the parameter. In the context of CMAB-T, a widespread smoothness relation allowing to treat objective functions resembling to linear functions is the following ℓ_1-norm triggering probability modulated condition (Wang and Chen, 2017):

$$\forall S \in \mathcal{S},\ \forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^n,\ \left|r(S;\boldsymbol{\mu}) - r(S;\boldsymbol{\mu}')\right| \leq B \sum_i p_i(S) |\mu_i - \mu_i'|,$$

where B is a constant. For the right choice of B, this relation holds in all the CMAB-T problems we considered previously. In the following, we give an example of bonus built from such relation for the reward function. To do so, we use a confidence region on the parameter vector together with the following Hölder's inequality for $p \geq 1$:

$$\sum_i p_i(S_t) \left|\overline{\mu}_{i,t-1} - \mu_i^*\right| \leq \left\| \left(\frac{p_i(S_t)}{N_{i,t-1}^{1/2}} \right)_i \right\|_p \underbrace{\left\| \left(N_{i,t-1}^{1/2} \left|\overline{\mu}_{i,t-1} - \mu_i^*\right| \right)_i \right\|_{\frac{p}{p-1}}}_{(3.3)}.$$

Here, the confidence region is defined as a bound on (3.3). For simplicity's sake, let's assume that outcomes received are Gaussian of unit variance, and that $N_{i,t-1}$ is not random (thus, the vector inside the $p/(p-1)$-norm have $\mathcal{N}(0,1)$ components). Furthermore, we consider two cases: arbitrary correlated and independent outcomes. For arbitrary correlated outcomes, we take $p = 1$ and get that (3.3) $= \mathcal{O}\left(\sqrt{\log(t)}\right)$ with high probability.[2] For independent outcomes, we take $p = 2$ so that (3.3) follows a χ distribution. Again, with high probability, we have (3.3) $\leq \sqrt{n + 2\sqrt{n\log(t)} + 2\log(t)} = \mathcal{O}\left(\sqrt{\log(t)}\right)$ (Laurent and Massart, 2000). In any case, (3.3) is bounded by $\mathcal{O}\left(\sqrt{\log(t)}\right)$ with high probability. To sum up, the bonus is an ℓ_p-norm of some error vector of the form "probability of observing i" (here, $p_i(S_t)$) times "error associated with i" (here, $c\sqrt{\log(t)/N_{i,t-1}}$ for some constant c). In what follows we consider a more general form for the error vector.

3.4.2 Several types of bonuses, several types of regret rates

In this subsection, we bound the regret for various types of bonuses. We rely on some results that we will state later in the "appendix" subsection 3.4.3.

Initialization As we will notice, in the propositions from subsection 3.4.3, the counters start at 1. There are several ways to deal with the case where some counters are 0. A first possibility is to use a few rounds in order to properly initialize all the counters, for example, for $t = i$, by choosing S_t such that $p_i(S_t) = 1$, we are assured that after n rounds, all the counters are at least equal to 1. Thus, such an initialization only adds the additive term $\sum_{i\in[n]} \Delta_{i,\max}$ to the regret. Another possibility is when

[2] Since each component is $\mathcal{O}(\sqrt{\log(t)})$, the ℓ_∞-norm is so.

an a priori bound on Δ_t of the form $\sum_{i\in[n]} p_i(S_t)K_i$ is known. This is the purpose of the following proposition.

Proposition 5 (Initialization)**.** *For all $i \in [n]$, let $K_i \in \mathbb{R}_+$. For all $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \sum_{i\in[n],\ N_{i,t-1}=0} p_i(S_t)K_i \right\}.$$

Then, if $\{t \leq T\} \in \mathcal{F}_t$, the event-filtered regret $\mathbb{E}\left[\sum_{t=1}^T \Delta_t \mathbb{I}\{\mathfrak{A}_t\}\right]$ is upper bounded by $\sum_{i\in[n]} K_i$.

Proof. Since $\mathbb{I}\{t \leq T\} \in \mathcal{F}_t$ and by definition of $p_i(S_t) = \mathbb{P}[i \in A_t|\mathcal{F}_t]$, we have

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right] \leq \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}\left[\sum_{i\in[n]} \mathbb{I}\{N_{i,t-1}=0,\ i\in A_t\}\middle|\mathcal{F}_t\right] K_i\right].$$

Since the counter $N_{i,t-1}$ is updated as soon as $i \in A_t$, we have that the event can occur for at most one round, giving the upper bound $\sum_{i\in[n]} K_i$. □

Analysis for ℓ_1-bonuses We give here an analysis in the case where the bonus is built from the ℓ_1-norm of some error vector. As we will see in Chapter 4, bonuses based on the ℓ_1-norm are obtained with a hypercube-type confidence region. This kind of confidence region is always possible to establish under the right concentration conditions for the marginals. Our formulation here is general enough to include many types of error vectors, and it will be easy to adapt each parameter to our needs. In the proof of the following Theorem 12, we use a technique called *reverse amortisation* (Wang and Chen, 2017), which replaces the analysis from (Kveton, Wen, Ashkan, and Szepesvari, 2015b). Reverse amortisation is used to transform a bound on Δ_t by considering only the error components that are large enough. It is based on the following observation: for a random variable Z such that $\mathbb{E}[Z] \geq 0$, we have $\mathbb{E}[Z] \leq \mathbb{E}[2Z\mathbb{I}\{2Z \geq \mathbb{E}[Z]\}]$. This is tight, since taking

$$Z = \begin{cases} 1/2 - \varepsilon & \text{with probability } 1-\varepsilon \\ \varepsilon^{-1}(1-(1/2-\varepsilon)(1-\varepsilon)) & \text{with probability } \varepsilon, \end{cases}$$

we have $\mathbb{E}[Z] = 1$ and $\mathbb{E}[2Z\mathbb{I}\{2Z \geq \mathbb{E}[Z]\}] = 1 + \mathcal{O}(\varepsilon)$. The fast reader can skip the proof of the following theorem to directly go to Theorem 13.

Theorem 12 (Regret bound for ℓ_1-bonus)**.** *For all $i \in [n]$, let $(\alpha_i, \beta_{i,T}) \in (0,1] \times \mathbb{R}_+$. For all $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \left\| \sum_{i\in[n],\ N_{i,t-1}>0} \frac{p_i(S_t)b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_1 \right\}.$$

Then, if $\{t \leq T\} \in \mathcal{F}_t$, the event-filtered regret $\mathbb{E}\left[\sum_{t=1}^T \Delta_t \mathbb{I}\{\mathfrak{A}_t\}\right]$ is upper bounded by

$$\sum_{i\in[n]} \mathbb{E}[\beta_{i,T}]\gamma_i \left(\mathbb{I}\{\alpha_i = 1\} 2\left(2 + \log\left(\frac{\Delta_{i,\max}}{\gamma_i \delta_{i,\min}}\right)\right) + \mathbb{I}\{\alpha_i < 1\} \frac{2^{1/\alpha_i}}{1-\alpha_i} {\delta_{i,\min}}^{1-1/\alpha_i} \right),$$

where

$$\gamma_i = \max_{S\in\mathcal{S},\ p_i(S)>0} b_i(S),$$

$$\delta_{i,\min} = \min_{S\in\mathcal{S},\ p_i(S)>0} \frac{\Delta(S)}{\sum_{j\in[n]} p_j(S)b_j(S)},$$
$$\delta_{i,\max} = \max_{S\in\mathcal{S},\ p_i(S)>0} \frac{\Delta(S)}{\sum_{j\in[n]} p_j(S)b_j(S)}.$$

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

$$\Delta_t \leq -\Delta_t + \left\| \sum_{i\in[n],\ N_{i,t-1}>0} \frac{2p_i(S_t)b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_1 \tag{3.4}$$
$$= -\frac{\sum_{i\in[n]} \Delta_t p_i(S_t)b_i(S_t)}{\sum_{i\in[n]} p_i(S_t)b_i(S_t)} + \sum_{i\in[n],\ N_{i,t-1}>0} \frac{2p_i(S_t)b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}}$$
$$\leq \sum_{i\in[n]} p_i(S_t)b_i(S_t) 0 \vee \left(\frac{2\mathbb{I}\{N_{i,t-1}>0\}\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Delta_t}{\sum_{j\in[n]} p_j(S_t)b_j(S_t)} \right) \tag{3.5}$$
$$\leq \sum_{i\in[n],\ N_{i,t-1}>0} \mathbb{I}\left\{ \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Delta_t}{\sum_{j\in[n]} p_j(S_t)b_j(S_t)} \right\} \frac{2p_i(S_t)\gamma_i\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \tag{3.6}$$

where (3.4) uses the event $\mathfrak{A}_t$, (3.5) uses the sub-additivity of $x \mapsto 0 \vee x$, (3.6) uses the fact that $0 \vee (x-y) \leq x\mathbb{I}\{x \geq y\}$ if $y \geq 0$ and $x \in \mathbb{R}$. Now, by definition of $p_i(S_t) = \mathbb{P}[i \in A_t | \mathcal{F}_t]$, we can get from (3.6) that Δ_t is upper bounded by

$$\sum_{i\in[n]} \mathbb{E}\left[\mathbb{I}\left\{ i \in A_t,\ p_i(S_t), \Delta_t, N_{i,t-1} > 0,\ \frac{2\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Delta_t}{\sum_{j\in[n]} p_j(S_t)b_j(S_t)} \right\} \frac{2\gamma_i\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \middle| \mathcal{F}_t \right].$$

Letting $f_i(x) = \beta_{i,T} 2^{1/\alpha_i} x^{-1/\alpha_i}$, the previous upper bound rewrites in the following way, with $\delta_t = \Delta_t / \sum_{j\in[n]} p_j(S_t)b_j(S_t)$:

$$\Delta_t \leq \sum_{i\in[n]} \mathbb{E}\Big[\gamma_i \mathbb{I}\{ i \in A_t,\ p_i(S_t), \Delta_t, N_{i,t-1} > 0,\ N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1}) \Big| \mathcal{F}_t \Big].$$

Now, we want to apply Proposition 8 and Proposition 9 from the next subsection 3.4.3. To do so, we distinguish two cases: $\alpha_i = 1$ and $\alpha_i < 1$.

$$\underbrace{\sum_{i\in[n]} \gamma_i \mathbb{I}\{\alpha_i = 1,\ i \in A_t,\ p_i(S_t), \Delta_t, N_{i,t-1} > 0,\ N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1})}_{(3.7)_t}$$
$$+ \underbrace{\sum_{i\in[n]} \gamma_i \mathbb{I}\{\alpha_i < 1,\ i \in A_t,\ p_i(S_t), \Delta_t, N_{i,t-1} > 0,\ N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1})}_{(3.8)_t}.$$

We consider the event

$$\mathfrak{B}_t \triangleq \left\{ \forall i \in A_t \text{ such that } \alpha_i = 1,\ p_i(S_t) > 0, \text{ we have } N_{i,t-1} > f_i\left(\frac{\Delta_{i,\max}}{\gamma_i} \right) \right\}$$

that allows us to write the following upper bound

$$\mathbb{I}\{\mathfrak{A}_t\}\Delta_t \leq \mathbb{E}[(3.8)_t + \mathbb{I}\{\mathfrak{B}_t\}(3.7)_t + \mathbb{I}\{\neg\mathfrak{B}_t\}\Delta_t|\mathcal{F}_t].$$

Since $\mathbb{I}\{t \leq T\} \in \mathcal{F}_t$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}[(3.8)_t + \mathbb{I}\{\mathfrak{B}_t\}(3.7)_t + \mathbb{I}\{\neg\mathfrak{B}_t\}\Delta_t|\mathcal{F}_t]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}(3.8)_t + \mathbb{I}\{\mathfrak{B}_t\}(3.7)_t + \mathbb{I}\{\neg\mathfrak{B}_t\}\Delta_t\right].\end{aligned}$$

Using Proposition 9, we handle the sum over t of the first term:

$$\mathbb{E}\left[\sum_{t=1}^{T}(3.8)_t\right] \leq \sum_{i\in[n]}\gamma_i\mathbb{I}\{\alpha_i < 1\}\frac{2^{1/\alpha_i}\mathbb{E}[\beta_{i,T}]}{1-\alpha_i}{\delta_{i,\min}}^{1-1/\alpha_i}.$$

Using Proposition 8, we can handle the sum over t of the second term:

$$\mathbb{E}\left[\sum_{t=1}^{T}(3.7)_t\mathbb{I}\{\mathfrak{B}_t\}\right] \leq \sum_{i\in[n]}\mathbb{I}\{\alpha_i = 1\}2\gamma_i\mathbb{E}[\beta_{i,T}]\left(1+\log\left(\frac{\Delta_{i,\max}}{\gamma_i\delta_{i,\min}}\right)\right).$$

The choice $\Delta_{i,\max}/\gamma_i$ is justified when we handle $\sum_{t=1}^{T}\Delta_t\mathbb{I}\{\neg\mathfrak{B}_t\}$:

$$\begin{aligned}\Delta_t\mathbb{I}\{\neg\mathfrak{B}_t\} &\leq \sum_{i\in A_t,\alpha_i=1}\mathbb{I}\left\{p_i(S_t) > 0,\ 0 < N_{i,t-1} \leq f_i\left(\frac{1}{\gamma_i}\Delta_{i,\max}\right)\right\}\Delta_t \\ &\leq \sum_{i\in A_t,\ \alpha_i=1}\mathbb{I}\left\{p_i(S_t) > 0,\ 0 < N_{i,t-1} \leq f_i\left(\frac{1}{\gamma_i}\Delta_{i,\max}\right)\right\}\Delta_{i,\max}.\end{aligned}$$

So, by summing over $t \in [T]$, we get that $\sum_{t=1}^{T}\Delta_t\mathbb{I}\{\neg\mathfrak{B}_t\}$ is upper bounded by

$$\begin{aligned}&\sum_{i\in[n],\ \alpha_i=1}\Delta_{i,\max}\left(\sum_{t=1}^{T}\mathbb{I}\left\{i \in A_t,\ 0 < N_{i,t-1} \leq f_i\left(\frac{1}{\gamma_i}\Delta_{i,\max}\right)\right\}\right) \\ &\leq \sum_{i\in[n]}\Delta_{i,\max}f_i\left(\frac{1}{\gamma_i}\Delta_{i,\max}\right)\mathbb{I}\{\alpha_i = 1\} \\ &= \sum_{i\in[n]}2\gamma_i\beta_{i,T}\mathbb{I}\{\alpha_i = 1\}.\end{aligned}$$

In summary, we have that $\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right]$ is upper bounded by

$$\sum_{i\in[n]}\mathbb{E}[\beta_{i,T}]\gamma_i\left(\mathbb{I}\{\alpha_i = 1\}2\left(2+\log\left(\frac{\Delta_{i,\max}}{\gamma_i\delta_{i,\min}}\right)\right) + \mathbb{I}\{\alpha_i < 1\}\frac{2^{1/\alpha_i}}{1-\alpha_i}{\delta_{i,\min}}^{1-1/\alpha_i}\right).$$

□

Analysis for ℓ_2-bonuses We now give the analysis in the case where the bonus is based on the ℓ_2-norm of the error vector. We will see in Chapter 5 that this kind of bonus appears naturally when the confidence region is ellipsoidal. In this case, there are several possibilities concerning the form of the bonus in question. Indeed,

contrary to the ℓ_1-norm, the ℓ_2-norm does not commute with the expectation (on the randomness of A_t). Here we give three different results according to the "position" of the expectation in the bonus. Each of the three results will find applications within the thesis.

The first result concerns the case where the expectation is within the norm. It generalizes the result provided in Degenne and Perchet (2016b) by incorporating it in the CMAB-T framework (their result was only stated within the CMAB setting). In addition, we provide here a simpler proof, like the simplification that Wang and Chen (2017) had proposed on the Kveton, Wen, Ashkan, and Szepesvari (2015b) method for ℓ_1 type bonuses.

Theorem 13 (Regret bound for ℓ_2-bonus, with expectation inside the norm)**.** *For all $i \in [n]$, let $(\alpha_i, \beta_{i,T}) \in [1/2, 1] \times \mathbb{R}_+$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \left\| \sum_{i \in [n],\ N_{i,t-1}>0} \frac{p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \right\}.$$

Then, if $\{t \leq T\} \in \mathcal{F}_t$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\} \Delta_t\right] \leq \sum_{i \in [n]} 4 \log_2(4\sqrt{m_i}) \mathbb{E}[\beta_{i,T}] \max_{S \in \mathcal{S},\ p_i(S)>0} b_i(S) \eta_i,$$

where

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

$$\delta_{i,\min} \triangleq \min_{S \in \mathcal{S},\ p_i(S)>0} \frac{\Delta(S)}{p_i(S) b_i(S)},$$

$$\delta_{i,\max} \triangleq \max_{S \in \mathcal{S},\ p_i(S)>0} \frac{\Delta(S)}{p_i(S) b_i(S)}.$$

Looking at the two previous theorems, we can see that there is a gain in using ℓ_2 type bonuses over ℓ_1 type bonuses. Indeed, in Theorem 13, note the disappearance of the term $\sum_{j \in [n]} p_j(S) b_j(S)$, which could potentially be quite large. For example, in the case where $A_t = S_t$ is of maximum cardinality m, $b_i(S_t) = 1$ and $\alpha_i = 1/2$, the use of a bonus of ℓ_2 improves the regret by a factor of $m \log^{-2}(m)$. Again, the following proof can be skipped to directly go to Theorem 14.

Proof of Theorem 13. Let $t \geq 1$, and define

$$\Lambda_t \triangleq \left\| \sum_{i \in [n]\ N_{i,t-1}>0} \frac{p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2,$$

such that the event $\mathfrak{A}_t$ rewrites as $\{\Delta_t \leq \Lambda_t\}$. We will only use this event at the end of the proof, and are focused on working with Λ_t for now. Our goal is first of all to bound Λ_t by an ℓ_2-norm where the components are controlled from the bottom and from the top. This then makes it possible to handle the ℓ_2-norm using a peeling

argument by grouping the components according to their range. We start by a simple lower bound on Λ_t, holding for any $j \in [n]$ such that $p_j(S_t) > 0$:

$$\Lambda_t \geq \left\| \frac{p_j(S_t) b_j(S_t) \beta_{j,T}^{\alpha_j} \mathbf{e}_j}{N_{j,t}^{\alpha_j}} \right\|_2 = \frac{p_j(S_t) b_j(S_t) \beta_{j,T}^{\alpha_j}}{N_{j,t}^{\alpha_j}}. \tag{3.9}$$

We then use a similar reverse amortisation technique than in the proof of Theorem 12, but this time with the ℓ_2-norm. For this, we define

$$m(S_t) \triangleq \sum_{i \in [n]} \mathbb{I}\{p_i(S_t) > 0\}$$

being the cardinality of the set of candidates to be in the triggering set A_t. We have that Λ_t is equal to

$$\begin{aligned}
&- \Lambda_t + 2 \left\| \sum_{i \in [n]} \sum_{N_{i,t-1} > 0} \frac{p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\
&= - \left\| \sum_{i \in [n]} \frac{\Lambda_t \mathbb{I}\{p_i(S_t) > 0\} \mathbf{e}_i}{\sqrt{m(S_t)}} \right\|_2 + \left\| \sum_{i \in [n]} \sum_{N_{i,t-1} > 0} \frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\
&\leq \left\| \sum_{i \in [n]} \mathbb{I}\{p_i(S_t), N_{i,t-1} > 0\} 0 \vee \left(\frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda_t}{\sqrt{m(S_t)}} \right) \mathbf{e}_i \right\|_2 \\
&= \left\| \sum_{i \in [n]} 0 \vee \left(\frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda_t}{\sqrt{m(S_t)}} \right) \mathbb{I}\{\mathfrak{B}_{i,t}\} \mathbf{e}_i \right\|_2 \\
&\leq \left\| \sum_{i \in [n]} \mathbb{I}\left\{ \mathfrak{B}_{i,t}, \frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda_t}{\sqrt{m(S_t)}} \right\} \frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i} \mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2,
\end{aligned}$$

where

$$\mathfrak{B}_{i,t} \triangleq \left\{ p_i(S_t), N_{i,t-1} > 0, \Lambda_t \geq \frac{p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \right\},$$

and where the penultimate relation uses (3.9). We now decompose the interval $[1/\sqrt{m(S_t)}, 2]$ using a peeling:

$$[1/\sqrt{m(S_t)}, 2] \subset \bigcup_{k=0}^{\lceil \log_2(\sqrt{m(S_t)}) \rceil} [2^{-k}, 2^{1-k}].$$

This induces a partition of the set of indices:

$$\begin{aligned}
&\mathbb{I}\left\{ i \in [n],\ p_i(S_t), N_{i,t-1} > 0,\ 2\Lambda_t \geq \frac{2 p_i(S_t) b_i(S_t) \beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda_t}{\sqrt{m(S_t)}} \right\} \\
&\subset \bigcup_{k=0}^{\lceil \log_2(\sqrt{m(S_t)}) \rceil} J_{k,t},
\end{aligned}$$

where for all integer $1 \leq k \leq \left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil$,

$$J_{k,t} \triangleq \left\{ i \in [n],\ p_i(S_t), N_{i,t-1} > 0,\ 2^{1-k}\Lambda_t \geq \frac{2p_i(S_t)b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq 2^{-k}\Lambda_t \right\}.$$

We can thus upper bound Λ_t^2 using this decomposition:

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

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

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{k=0}^{\left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\}\Delta_t \\
&\leq \sum_{t=1}^{T} \sum_{k=0}^{\left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil} \mathbb{I}\{k_t = k,\ \mathfrak{A}_t\} \sum_{i\in[n]} \mathbb{I}\{i \in J_{k,t}\} 2^{2-2k}\left(\left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil + 1\right)\Delta_t \\
&= \sum_{i\in[n]} \sum_{t=1}^{T} \sum_{k=0}^{\left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil} \underbrace{\mathbb{I}\{k_t = k,\ \mathfrak{A}_t,\ i \in J_{k,t}\}}_{(3.10)_{i,k,t}} 2^{2-2k}\left(\left\lceil \log_2\left(\sqrt{m(S_t)}\right)\right\rceil + 1\right)\Delta_t.
\end{aligned} \tag{3.11}$$

Using the event $\mathfrak{A}_t$ and that $i \in J_{k,t}$, we have

$$\begin{aligned}
(3.10)_{i,k,t} &\leq \mathbb{I}\left\{ p_i(S_t), N_{i,t-1} > 0,\ N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+1}p_i(S_t)b_i(S_t)\beta_{i,T}^{\alpha_i}}{\Delta_t} \right\} \\
&= \underbrace{\mathbb{I}\left\{ p_i(S_t), N_{i,t-1} > 0,\ N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\delta_t} \right\}}_{(3.12)_{i,k,t}},
\end{aligned}$$

with

$$\delta_t = \Delta_t / (p_i(S_t)b_i(S_t)).$$

We now want to apply Proposition 7 from the next subsection 3.4.3. For this, we first need to add the event $\{i \in A_t\}$ to the previous indicator. As in Theorem 12, this will be done using $p_i(S_t)$. Then, we want to invert the sum over t and the one over k. We

thus have that (3.11) is bounded by

$$\sum_{i\in[n]}\sum_{t=1}^{T}\sum_{k=0}^{\lceil\log_2(\sqrt{m(S_t)})\rceil} p_i(S_t)(3.12)_{i,k,t}\; 2^{2-2k}\left(\left\lceil\log_2\left(\sqrt{m(S_t)}\right)\right\rceil+1\right)b_i(S_t)\delta_t \quad (3.13)$$

$$\leq \sum_{i\in[n]}\sum_{k=0}^{\lceil\log_2(\sqrt{m_i})\rceil}\sum_{t=1}^{T} p_i(S_t)(3.12)_{i,k,t}\; 2^{2-2k}(\lceil\log_2(\sqrt{m_i})\rceil+1)\max_{S\in\mathcal{S},\; p_i(S)>0} b_i(S)\delta_t$$

$$= \sum_{i\in[n]}\sum_{k=0}^{\lceil\log_2(\sqrt{m_i})\rceil} 2^{2-2k}(\lceil\log_2(\sqrt{m_i})\rceil+1)\max_{S\in\mathcal{S},\; p_i(S)>0} b_i(S)\underbrace{\sum_{t=1}^{T} p_i(S_t)(3.12)_{i,k,t}\delta_t}_{(3.14)_{i,k,t}}.$$

Using the definition of $p_i(S_t)$ and that $\{t\leq T\}\in\mathcal{F}_t$, we can write

$$\begin{aligned}
\mathbb{E}\Big[(3.14)_{i,k,t}\Big] &= \mathbb{E}\left[\sum_{t=1}^{T} p_i(S_t)\mathbb{I}\left\{p_i(S_t), N_{i,t-1}>0,\; N_{i,t-1}^{\alpha_i}\leq\frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\delta_t}\right\}\delta_t\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{I}\left\{i\in A_t,\; p_i(S_t), N_{i,t-1}>0,\; N_{i,t-1}^{\alpha_i}\leq\frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\delta_t}\right\}\delta_t\middle|\mathcal{F}_t\right]\right]\\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{i\in A_t,\; p_i(S_t), N_{i,t-1}>0,\; N_{i,t-1}^{\alpha_i}\leq\frac{2^{k+1}\beta_{i,T}^{\alpha_i}}{\delta_t}\right\}\delta_t\right]}_{(3.15)_{i,k}}
\end{aligned}$$

Now, Proposition 7 gives

$$(3.15)_{i,k}\leq\mathbb{E}[\beta_{i,T}]\left(\mathbb{I}\{\alpha_i<1\}\frac{2^{\frac{k+1}{\alpha_i}}}{1-\alpha_i}\delta_{i,\min}^{1-1/\alpha_i}+\mathbb{I}\{\alpha_i=1\}2^{k+1}\left(1+\log\left(\frac{\delta_{i,\max}}{\delta_{i,\min}}\right)\right)\right).$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right]\leq\sum_{i\in[n]}4\log_2(4\sqrt{m_i})\mathbb{E}[\beta_{i,T}]\max_{S\in\mathcal{S},\; p_i(S)>0} b_i(S)\eta_i,$$

where

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

□

The second result we propose starts from the previous bonus (the one provided in Theorem 13) and uses Jensen's inequality in order to bound it by placing the expectation outside the norm. One can thus see that the result given in the following Theorem 14 can be used instead of the previous one. We can nevertheless notice that the result provided is a little less good. For example, when $\alpha_i=1/2$, and $b_i(S_t)=b$, $p_i(S_t)=p$, $\Delta(S_t)=\Delta$, we see that the factor $p\in[0,1]$ disappears in the bound of

Theorem 14, compared to the one from Theorem 13. We can skip the proof of the following theorem to go to Theorem 15.

Theorem 14 (Regret bound for ℓ_2-bonus, with expectation outside the norm)**.** *For all $i \in [n]$, let $(\alpha_i, \beta_{i,T}) \in [1/2, 1) \times \mathbb{R}_+$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{ \Delta_t \leq \mathbb{E}\left[\left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \middle| \mathcal{F}_t \right] \right\}.$$

Then, if $\{t \leq T\} \in \mathcal{F}_t$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right] \leq \sum_{i \in [n]} 4\log_2(4\sqrt{m_i}) \max_{S \in \mathcal{S},\ p_i(S)>0} b_i(S)^{\frac{1}{\alpha_i}} \mathbb{E}[\beta_{i,T}]\eta_i,$$

where

$$\eta_i = \begin{cases} 32\log_2(4\sqrt{m_i})\Delta_{i,\min}^{-1} & \text{if } \alpha_i = 1/2 \\ 2^{\frac{2}{\alpha_i}}\left(\left(1 - 2^{\frac{1}{\alpha_i}-2}\right)(1-\alpha_i)\Delta_{i,\min}^{\frac{1-\alpha_i}{\alpha_i}}\right)^{-1} & \text{if } 1/2 < \alpha_i < 1. \end{cases}$$

Proof. Let $t \geq 1$. With a first reverse amortisation, we start by restricting the set of possibles for A_t by only taking those whose error is at least twice as large as Δ_t: assuming that $\mathfrak{A}_t$ holds, we have

$$\begin{aligned} \Delta_t &\leq \mathbb{E}\left[2\left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 - \Delta_t \middle| \mathcal{F}_t \right] \\ &\leq \mathbb{E}\left[\mathbb{I}\left\{ \left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{2b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \geq \Delta_t \right\} \left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{2b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \middle| \mathcal{F}_t \right] \end{aligned}$$

We now define

$$\Lambda(A_t) \triangleq \left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{2b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2,$$

and have for any $j \in A_t$ that

$$\Lambda(A_t) \geq \frac{2b_j(S_t)\beta_{j,T}^{\alpha_j}}{N_{j,t}^{\alpha_j}}. \tag{3.16}$$

Then, a similar technique as in Theorem 13 gives that $\Lambda(A_t)$ equals

$$\begin{aligned} &-\Lambda(A_t) + \left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\ &= -\left\| \sum_{i \in A_t} \frac{\Lambda(A_t)\mathbf{e}_i}{\|\mathbf{e}_{A_t}\|_2} \right\|_2 + \left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2 \\ &\leq \left\| \sum_{i \in A_t, N_{i,t-1}>0} 0 \vee \left(\frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda(A_t)}{\|\mathbf{e}_{A_t}\|_2} \right) \mathbf{e}_i \right\|_2 \end{aligned}$$

$$= \left\| \sum_{i \in A_t, N_{i,t-1}>0} 0 \vee \left(\frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} - \frac{\Lambda(A_t)}{\|\mathbf{e}_{A_t}\|_2} \right) \mathbb{I}\left\{ \Lambda(A_t) \geq \frac{2b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \right\} \mathbf{e}_i \right\|_2 \tag{3.16}$$

$$\leq \left\| \sum_{i \in A_t, N_{i,t-1}>0} \mathbb{I}\left\{ 2\Lambda(A_t) \geq \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda(A_t)}{\|\mathbf{e}_{A_t}\|_2} \right\} \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2.$$

We now consider the following partition of the set of indices:

$$\mathbb{I}\left\{ i \in A_t, N_{i,t-1} > 0,\ 2\Lambda(A_t) \geq \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda(A_t)}{\|\mathbf{e}_{A_t}\|_2} \right\} \subset \bigcup_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} J_{k,t},$$

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

$$J_{k,t} \triangleq \left\{ i \in A_t, N_{i,t-1} > 0,\ 2^{1-k}\Lambda(A_t) \geq \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq 2^{-k}\Lambda(A_t) \right\}.$$

We bound $\Lambda(A_t)^2$ as

$$\begin{aligned}
\Lambda(A_t)^2 &\leq \left\| \sum_{i \in A_t, N_{i,t-1}>0} \mathbb{I}\left\{ 2\Lambda(A_t) \geq \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}}{N_{i,t-1}^{\alpha_i}} \geq \frac{\Lambda(A_t)}{\|\mathbf{e}_{A_t}\|_2} \right\} \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2^2 \\
&= \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} \left\| \sum_{i \in J_{k,t}} \frac{4b_i(S_t)\beta_{i,T}^{\alpha_i}\mathbf{e}_i}{N_{i,t-1}^{\alpha_i}} \right\|_2^2 \\
&\leq \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} 2^{2-2k}\Lambda(A_t)^2 \left\| \mathbf{e}_{J_{k,t}} \right\|_2^2.
\end{aligned}$$

So there is an integer k_t such that $|J_{k_t,t}| = \left\| \mathbf{e}_{J_{k_t,t}} \right\|_2^2 \geq 2^{2k_t-2}(1 + \lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil)^{-1}$.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\}\Delta_t \leq \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}\{\Lambda(A_t) \geq \Delta_t\}\Lambda(A_t)|\mathcal{F}_t] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} \mathbb{I}\{k_t = k,\ \Lambda(A_t) \geq \Delta_t\}\Lambda(A_t) \middle| \mathcal{F}_t \right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} \mathbb{I}\{k_t = k,\ \Lambda(A_t) \geq \Delta_t\} \frac{\sum_{i \in [n]} \mathbb{I}\{i \in J_{k,t}\}}{2^{2k-2}(1 + \lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil)^{-1}} \Lambda(A_t) \middle| \mathcal{F}_t \right] \\
&\leq \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil} \frac{\mathbb{I}\left\{ i \in A_t, 0 < N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+2}b_i(S_t)\beta_{i,T}^{\alpha_i}}{\Lambda(A_t)}, \Lambda(A_t) \geq \Delta_t \right\}}{2^{2k-2}(1 + \lceil \log_2(\|\mathbf{e}_{A_t}\|_2) \rceil)^{-1}} \Lambda(A_t) \middle| \mathcal{F}_t \right].
\end{aligned}$$

Taking the expectation of the above, and using $\{t \leq T\} \in \mathcal{F}_t$, we have the bound

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\}\Delta_t \right]$$

$$\leq \sum_{i=1}^{n} \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=0}^{\lceil \log_2(\|\mathbf{e}_{A_t}\|_2)\rceil} \frac{\mathbb{I}\left\{i \in A_t,\ 0 < N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+2} b_i(S_t) \beta_{i,T}^{\alpha_i}}{\Lambda(A_t)},\ \Lambda(A_t) \geq \Delta_t\right\}}{2^{2k-2}(1+\lceil \log_2(\|\mathbf{e}_{A_t}\|_2)\rceil)^{-1}} \Lambda(A_t)\right]$$

$$\leq \sum_{i=1}^{n} \sum_{k=0}^{\lceil \log_2(\sqrt{m_i})\rceil} \mathbb{E}\left[\frac{1+\lceil \log_2(\sqrt{m_i})\rceil}{2^{2k-2}} (3.17)_{i,k}\right],$$

where

$$(3.17)_{i,k} \triangleq \sum_{t=1}^{T} \mathbb{I}\left\{i \in A_t,\ 0 < N_{i,t-1}^{\alpha_i} \leq \frac{2^{k+2} b_i(S_t) \beta_{i,T}^{\alpha_i}}{\Lambda(A_t)},\ \Lambda(A_t) \geq \Delta_t\right\} \Lambda(A_t).$$

Applying Proposition 7 from the next subsection 3.4.3 gives

$$(3.17)_{i,k} \leq \frac{\max_{S \in \mathcal{S},\ p_i(S)>0} b_i(S)^{\frac{1}{\alpha_i}} \beta_{i,T} 2^{\frac{k+2}{\alpha_i}}}{1-\alpha_i} \Delta_{i,\min}^{1-1/\alpha_i},$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right] \leq \sum_{i \in [n]} 4 \log_2(4\sqrt{m_i}) \max_{S \in \mathcal{S},\ p_i(S)>0} b_i(S)^{\frac{1}{\alpha_i}} \mathbb{E}[\beta_{i,T}] \eta_i,$$

where

$$\eta_i = \begin{cases} 32 \log_2(4\sqrt{m_i}) \Delta_{i,\min}^{-1} & \text{if } \alpha_i = 1/2 \\ 2^{\frac{2}{\alpha_i}} \left(\left(1 - 2^{\frac{1}{\alpha_i}-2}\right)(1-\alpha_i) \Delta_{i,\min}^{\frac{1-\alpha_i}{\alpha_i}}\right)^{-1} & \text{if } 1/2 < \alpha_i < 1. \end{cases}$$

□

Finally, the last result re-uses Jensen's inequality in the bonus from Theorem 14, with the concavity of the square root. Again, the result loses power compared to the Theorems 13, 14. In fact, we see that by linearity of the ℓ_2-norm squared, we are reduced to a particular form of Theorem 13 where, in the error vector, the probability $p_i(S_t)$ is changed to the square root of this probability.

Theorem 15 (Regret bound for ℓ_2-bonus, with expectation between the root and the squared norm)**.** *For all $i \in [n]$, let $\beta_{i,T} \in \mathbb{R}_+$. For $t \geq 1$, consider the event*

$$\mathfrak{A}_t \triangleq \left\{\Delta_t \leq \sqrt{\mathbb{E}\left[\left\| \sum_{i \in A_t, N_{i,t-1}>0} \frac{b_i(S_t) \beta_{i,T}^{1/2} \mathbf{e}_i}{N_{i,t-1}^{1/2}} \right\|_2^2 \middle| \mathcal{F}_t\right]}\right\}.$$

Then, if $\{t \leq T\} \in \mathcal{F}_t$, we have that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right]$ is bounded by

$$\sum_{i \in [n]} 16 \log_2^2(4\sqrt{m_i}) \mathbb{E}[\beta_{i,T}] \max_{S \in \mathcal{S},\ p_i(S)>0} \frac{b_i(S)^2}{\Delta(S)} (1 + \log(\eta_i)),$$

where

$$\eta_i \triangleq \frac{\max_{S \in \mathcal{S},\ p_i(S)>0} \Delta(S)^2 / (p_i(S) b_i(S)^2)}{\min_{S \in \mathcal{S},\ p_i(S)>0} \Delta(S)^2 / (p_i(S) b_i(S)^2)}.$$

Proof. Let $t \geq 1$ and notice that

$$\sqrt{\mathbb{E}\left[\left\|\sum_{i\in A_t, N_{i,t-1}>0} \frac{b_i(S_t)\beta_{i,T}^{1/2}\mathbf{e}_i}{N_{i,t-1}^{1/2}}\right\|_2^2 \middle| \mathcal{F}_t\right]} = \left\|\sum_{i\in[n], N_{i,t-1}>0} \frac{\sqrt{p_i(S_t)}b_i(S_t)\beta_{i,T}^{1/2}\mathbf{e}_i}{N_{i,t-1}^{1/2}}\right\|_2.$$

So, we can use Theorem 13 with $\alpha_i = 1/2$ and another b_i function, that we denote $\widetilde{b}_i$, and that is defined as

$$\widetilde{b}_i(S) = b_i(S)\left(\frac{1}{\sqrt{p_i(S)}}\mathbb{I}\{p_i(S) > 0\} + \mathbb{I}\{p_i(S) = 0\}\right).$$

We however do not use the result directly, but rather rework the end of the proof in order to get a logarithmic dependence in p_i. More precisely, the proof remains the same until the relation (3.13). In particular, we have the definition

$$\delta_t = \Delta_t/(p_i(S_t)\widetilde{b}_i(S_t)).$$

Then, in (3.13), we bound

$$\mathbb{I}\{p_i(S_t) > 0\}\widetilde{b}_i(S_t)\delta_t = \mathbb{I}\{p_i(S_t) > 0\}\frac{\Delta_t}{p_i(S_t)}$$

by

$$\max_{S\in\mathcal{S},\ p_i(S)>0} \frac{b_i(S)^2}{\Delta(S)}\mathbb{I}\{p_i(S_t) > 0\}\frac{\Delta_t^2}{p_i(S_t)b_i(S_t)^2} = \max_{S\in\mathcal{S},\ p_i(S)>0} \frac{b_i(S)^2}{\Delta(S)}\mathbb{I}\{p_i(S_t) > 0\}\delta_t^2.$$

The end of the proof is the same, except from the application of Proposition 7, where having δ_t^2 instead of δ_t gives a behavior as if $\alpha_i = 1$, giving the final bound

$$\sum_{i\in[n]} 16\log_2^2(4\sqrt{m_i})\mathbb{E}[\beta_{i,T}] \max_{S\in\mathcal{S},\ p_i(S)>0} \frac{b_i(S)^2}{\Delta(S)}\left(1 + \log\left(\frac{\delta_{i,\max}^2}{\delta_{i,\min}^2}\right)\right).$$

□

The regret bound given in Theorem 15 has a logarithmic dependence in p_i. A question that arises then is: can we express a bound that does not depend at all on p_i? The answer is yes, and it is done using Proposition 9 instead of Proposition 7 at the end of the proof. This is the purpose of the following Theorem 16.

Theorem 16. *Under the same assumptions as in Theorem 15, we have the following upper bound on* $\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{\mathfrak{A}_t\}\Delta_t\right]$,

$$\sum_{i\in[n]} 16\log_2^2(4\sqrt{m_i})\mathbb{E}[\beta_{i,T}]\eta_i\left(1 + \mathbb{E}\left[\log\left(\frac{16\beta_{i,T}m_i}{\min_{S\in\mathcal{S},\ p_i(S)>0}\Delta(S)^2/b_i(S)^2}\right)\right]\right),$$

where

$$\eta_i = \max_{S\in\mathcal{S},\ p_i(S)>0} \frac{b_i(S)^2}{\Delta(S)}.$$

Proof. We use the same proof as in Theorem 15 until before the application of Proposition 7. In particular, we thus follow the beginning of the proof of Theorem 13. Before

applying Proposition 9 instead of Proposition 7, we first upper bound δ_t^2 by

$$\frac{2^{2k+2}\beta_{i,T}}{N_{i,t-1}},$$

Thanks to $(3.12)_{i,k,t}$. We then use Proposition 9 from the next subsection 3.4.3 with $\alpha_i = 1$, giving the logarithmic factor

$$1 + \log\left(\frac{2^{2k+2}\beta_{i,T}}{\delta_{i,\min}^2}\right)$$

instead of

$$1 + \log\left(\frac{\delta_{i,\max}^2}{\delta_{i,\min}^2}\right).$$

Finally, the desired result is obtained by bounding k by $\lceil \log_2(\sqrt{m_i}) \rceil \leq \log_2(2\sqrt{m_i})$, and by bounding $\delta_{i,\min}^{-2}$ by

$$\max_{S \in \mathcal{S},\ p_i(S)>0} b_i(S)^2/\Delta(S)^2.$$

□

When we derive an upper bound on Δ_t, it may happen that we do not fall into one of the cases mentioned in the above theorems. Specifically, it may be that the bonus is composed of several terms, each of which has the proper form to be treated as above. In addition, when initialization is handled with Proposition 5, this adds an additional term (corresponding to zero counters), so it is useful to know how to handle it. We will see in the following Proposition 6 that previous results are sufficient to treat a composed bonus.

Proposition 6 (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 is a 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., $\Delta_t \leq \sum_{k \in [K]} \Delta_t \mathbb{I}\{\Delta_t \leq K B_{k,t}\}$. □

We can therefore see that we can use Proposition 6 coupled with the above theorems in order to treat the composed bonuses. Doing so, we only lose a factor K in the previous analyses. Typically, K is a small universal constant, so we do not lose in the regret rate.

3.4.3 Appendix

As we have already seen in Theorem 10 for MAB, once we had managed to bound the gap Δ_t by the bonus, a rearrangement of the inequality made it possible to have that

the counter of the chosen arm is bounded by a function of the minimal gap, leading to the final bound by summing over t. We will see here that the same principle also applies in the CMAB-T context. As we saw, this is due to the fact that the bonuses are each time expressed using the counters $N_{i,t-1}$ for $i \in A_t$. We give in this subsection a series of results that, for an arm i, bound the regret filtered by the event where $i \in A_t$ and an upper bound on a counter $N_{i,t-1}$ is known. These results are extending those given in Chen, Wang, and Yuan (2013), Chen, Wang, and Yuan (2016), and Wang and Chen (2017).

Proposition 7. *Let $i \in [n]$ and $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing function, integrable on an interval $[\delta_{i,\min}, \delta_{i,\max}] \subset \mathbb{R}_+^*$. Then for any sequence of real numbers $(\delta_t) \in ([\delta_{i,\min}, \delta_{i,\max}] \cup \{0\})^T$,*

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

In particular,

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

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

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

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

Proof. Consider $\delta_{i,\max} = \delta_{i,1} \geq \delta_{i,2} \geq \cdots \geq \delta_{i,K_i} = \delta_{i,\min}$ being all possible values for δ_t when $\delta_t \neq 0$. We define a dummy gap $\delta_{i,0} = \infty$ and let $f_i(\delta_{i,0}) = 0$. In (3.18), we look at times t where $\delta_t \neq 0$ and 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 $\delta_t \neq 0$. Notice that in (3.18), we do not explicitly use k_t, but instead sum over all $k \in [K_i]$ and filter against the event $\{\delta_{i,k} \geq \delta_t\}$, which is equivalent to summing over $k \in [k_t]$.

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ N_{i,t-1} \leq f_i(\delta_t)\}\delta_t \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\delta_{i,k-1}) < N_{i,t-1} \leq f_i(\delta_{i,k}), \delta_{i,k} \geq \delta_t\}\delta_t. \qquad (3.18)\end{aligned}$$

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 gap δ_t by $\delta_{i,k}$.

$$(3.18) \leq \sum_{t=1}^{T} \sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\delta_{i,k-1}) < N_{i,t-1} \leq f_i(\delta_{i,k}), \delta_{i,k} \geq \delta_t\}\delta_{i,k}. \tag{3.19}$$

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 $\{\delta_{i,k} \geq \delta_t\}$.

$$(3.19) \leq \sum_{t=1}^{T} \sum_{k=1}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\delta_{i,k-1}) < N_{i,t-1} \leq f_i(\delta_{i,k})\}\delta_{i,k}. \tag{3.20}$$

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

$$(3.20) = \sum_{k=1}^{K_i} \sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ f_i(\delta_{i,k-1}) < N_{i,t-1} \leq f_i(\delta_{i,k})\}\delta_{i,k}. \tag{3.21}$$

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 $\lfloor y \rfloor - \lfloor x \rfloor$.

$$(3.21) \leq \sum_{k=1}^{K_i} (\lfloor f_i(\delta_{i,k}) \rfloor - \lfloor f_i(\delta_{i,k-1}) \rfloor)\delta_{i,k}. \tag{3.22}$$

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

$$(3.22) = \lfloor f_i(\delta_{i,K_i}) \rfloor \delta_{i,K_i} + \sum_{k=1}^{K_i - 1} \lfloor f_i(\delta_{i,k}) \rfloor (\delta_{i,k} - \delta_{i,k+1}) \tag{3.23}$$

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 $k \in [K_i - 1]$.

$$(3.23) \leq f_i(\delta_{i,K_i})\delta_{i,K_i} + \sum_{k=1}^{K_i - 1} f_i(\delta_{i,k})(\delta_{i,k} - \delta_{i,k+1}) \tag{3.24}$$

$$\leq f_i(\delta_{i,K_i})\delta_{i,K_i} + \int_{\delta_{i,K_i}}^{\delta_{i,1}} f_i(x)\mathrm{d}x. \tag{3.25}$$

□

Proposition 8. *Let $i \in [n]$ and $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a decreasing function, integrable on $[\delta_{i,\min}, \delta_{i,\max}] \subset \mathbb{R}_+^*$. Then for any sequence of real numbers $(\delta_t) \in$*

$([\delta_{i,\min}, \delta_{i,\max}] \cup \{0\})^T$,

$$\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ \delta_t \neq 0,\ f_i(\delta_{i,\max}) < N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1})$$
$$\leq f_i(\delta_{i,\min})\delta_{i,\min} + \int_{\delta_{i,\min}}^{\delta_{i,\max}} f_i(x)\mathrm{d}x.$$

Proof. The proof is very similar to the one for Proposition 7. The key difference will be that instead of a right Riemann sum, this is a left Riemann sum that appears at the end of the proof. Therefore, we can't bound this Riemann sum directly by the integral. However, it can be made as close as desired to the integral by making the subdivision fine enough. This is possible since we never use any assumption on the subdivision of $[\delta_{i,\min}, \delta_{i,\max}]$ we consider (except that it contains all possible values for δ_t when $i \in A_t$, which can only reinforce its refinement). In addition, at the end, the subdivision only appears within the Riemann sum. We thus consider a general subdivision $\delta_{i,\max} = \delta_{i,1} \geq \delta_{i,2} \geq \dots \delta_{i,K_i} = \delta_{i,\min}$, that contains all possible values for δ_t when $\delta_t \neq 0$. Using this decomposition, we can write as previously

$$\sum_{t=1}^{T} \mathbb{I}\{i \in A_t,\ \delta_t \neq 0,\ f_i(\delta_{i,\max}) < N_{i,t-1} \leq f_i(\delta_t)\} f_i^{-1}(N_{i,t-1})$$
$$= \sum_{t=1}^{T} \sum_{k=2}^{K_i} \mathbb{I}\{i \in A_t,\ f_i(\delta_{i,k-1}) < N_{i,t-1} \leq f_i(\delta_{i,k}), \delta_{i,k} \geq \delta_t > 0\} f_i^{-1}(N_{i,t-1}). \quad (3.26)$$

Now, we can upper bound $f_i^{-1}(N_{i,t-1})$ by $\delta_{i,k-1}$ to recover a summation similar to (3.19) in the proof of Proposition 7. Using the same derivations as in (3.19)-(3.22), we get the upper bound

$$\begin{aligned}(3.26) &\leq \sum_{k=2}^{K_i} (\lfloor f_i(\delta_{i,k}) \rfloor - \lfloor f_i(\delta_{i,k-1}) \rfloor)\delta_{i,k-1} \\ &= \lfloor f_i(\delta_{i,K_i}) \rfloor \delta_{i,K_i} - \lfloor f_i(\delta_{i,1}) \rfloor \delta_{i,1} + \sum_{k=2}^{K_i} \lfloor f_i(\delta_{i,k}) \rfloor (\delta_{i,k-1} - \delta_{i,k}) \\ &\leq f_i(\delta_{i,K_i})\delta_{i,K_i} + \sum_{k=2}^{K_i} f_i(\delta_{i,k})(\delta_{i,k-1} - \delta_{i,k}),\end{aligned}$$

where in the last inequality, we first upper bound the term $-\lfloor f_i(\delta_{i,1}) \rfloor \delta_{i,1}$ by 0, and then use $\lfloor x \rfloor \leq x$ for all $x \in \mathbb{R}$. Finally, we recognize a left Riemann sum. We get the desired result taking the infimum over all possible subdivisions containing all possible values for δ_t when $\delta_t \neq 0$. □

Remark 9. *Proposition 8 does not exactly generalizes Proposition 7 since we have to filter against the event* $\{f_i(\delta_{i,\max}) < N_{i,t-1}\}$. *However, for* $f_i(x) = \beta_{i,T} x^{-1/\alpha_i}$, $\beta_{i,T} \geq 0$ *and* $\alpha_i \in (0,1)$, *the upper bound is*

$$\delta_{i,\min}^{1-1/\alpha_i} \frac{\beta_{i,T}}{1-\alpha_i} - \delta_{i,\max}^{1-1/\alpha_i} \frac{\alpha_i \beta_{i,T}}{1-\alpha_i} \leq \delta_{i,\min}^{1-1/\alpha_i} \frac{\beta_{i,T}}{1-\alpha_i},$$

so $\delta_{i,\max}$ *can be taken sufficiently large without altering the desired upper bound. In particular, for* $\delta_{i,\max}$ *such that* $f_i(\delta_{i,\max}) < 1$, *the event* $\{f_i(\delta_{i,\max}) < N_{i,t-1}\}$ *always*

holds (when the counter is non-zero). We will see another (simpler) proof of this fact in the next Proposition 9.

Proposition 9. *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*

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

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.

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

□

Chapter 4

An Example of CMAB-T Problem: Sequential Search-and-Stop

This chapter is based on Perrault, Perchet, and Valko (2019b), and aims to introduce a new example of a problem falling within the CMAB-T framework.

We consider here the problem where the agent wants to find a hidden object that is randomly located in some vertex of a directed acyclic graph (DAG) according to a fixed but possibly unknown distribution. The agent can only examine vertices whose in-neighbors have already been examined. We address a *learning* setting where we allow the agent to stop before having found the object and restart searching on a new independent instance of the same problem. Our goal is to maximize the total number of hidden objects found given a time budget. The agent can thus skip an instance after realizing that it would spend too much time on it. Our contributions are both to the *search theory* and *multi-armed bandits*. If the distribution is known, we provide a quasi-optimal and efficient stationary strategy. If the distribution is unknown, we additionally show how to sequentially approximate it and, at the same time, act near-optimally in order to collect as many hidden objects as possible.

4.1 Problem formulation and motivation

We study the setting where an object, called hider, is randomly located in one vertex of a directed acyclic graph (DAG), and where an agent wants to find it by sequentially selecting vertices one by one, and examining them at a (possibly random) cost. The agent has a strong constraint: its search must respect *precedence constraints* imposed by the DAG, i.e., a vertex can be examined only if *all* its in-neighbors have already been examined. The goal of the agent is to minimize the expected total search cost incurred before finding the hider. This setting is a type of *single machine scheduling* problem (Lín, 2015), where a set of n jobs $[n]$ have to be processed on a single machine that can process at most one job at a time. Once a job processing is started, it must continue without interruption until the processing is complete. Each job j has a cost c_j representing its processing time, and a weight w_j representing its importance. In our context, w_j is the probability that j contains the hider. The aim is to find a schedule (i.e., a permutation of jobs) that minimizes the total weighted completion time while respecting precedence constraints.[1] The setting was already shown to be NP-hard (Lawler, 1978; Lenstra and Rinnooy Kan, 1978). On the positive side, several polynomial-time α-approximations exist, depending on the assumption we take on the DAG (see e.g., the recent survey of Prot and Bellenguez-Morineau, 2017).

[1] The standard scheduling notation (Graham et al., 1979) denotes this setting as $1|prec|\sum w_j C_j$.

For instance, the case of $\alpha = 2$ can be dealt without any additional assumption. On the other hand, there is an exact $\mathcal{O}(n \log n)$-time algorithm when the partially ordered set (poset) defined by the DAG is a series-parallel order (Lawler, 1978). More generally, when the poset has fractional dimension of at most f, there is a polynomial-time approximation with $\alpha = 2 - 2/f$ (Ambühl, Mastrolilli, et al., 2011). In this work, we assume the DAG is such that an exact polynomial-time algorithm is available.[2] We denote this algorithm as SCHEDULING. For example, this is true for two-dimensional posets (Ambühl and Mastrolilli, 2009).

The problem is also well known in *search theory* (Stone, 1976; Fokkink, Lidbetter, and Végh, 2016), one of the disciplines originating from *operations research*. Since in our case, the search space is a DAG, we fall within the network search setting (Kikuta and Ruckle, 1994; Gal, 2001; Evans and Bishop, 2013). When the DAG is an out-tree, the problem reduces to the *expanding search* problem introduced by Alpern and Lidbetter (2013).

The case of unknown distribution of the hider is usually studied within the field of *search games*, i.e., a zero-sum game where the agent picks the search and plays against the hider with search cost as payoff (Alpern and Gal, 2006; Alpern, Fokkink, et al., 2013; Hohzaki, 2016). In our work, we deal with an unknown hider distribution by extending the stochastic setting to the sequential case, where at each round t, the agent faces a new, independent instance of the problem. The challenge is the need to learn the distribution through repeated interactions with the environment. Each instance, the agent has to perform a search based on the instances observed during the previous rounds. Furthermore, contrary to the typical search setting, the agent can additionally decide whether it wishes to abandon the search on the current instance and start a new one in the next round, even if the hider was not found. The goal of the agent is to collect as many hiders as possible, using a fixed budget B. This may be particularly useful, when the remaining vertices have large costs and it would not be cost-effective to examine them.

As a result, the hider may not be found in each round and the agent has to make a trade-off between exhaustive searches, which lead to a good estimation (*exploration*) and efficient searches, which leads to a good benefit/cost ratio (*exploitation*). The sequential *exploration-exploitation* trade-off is well studied in multi-armed bandits (Cesa-Bianchi and Lugosi, 2006; Lattimore and Szepesvári, 2019) and has been applied (as we saw in Chapter 2) to many fields including mechanism design (Mohri and Munoz, 2014), search advertising (Tran-Thanh, Stein, et al., 2014) and personalized recommendation (Li, Chu, et al., 2010). Since several vertices can be visited within each round, our setting can be seen as an instance of stochastic combinatorial semi-bandits. For this reason, we refer to a vertex $j \in [n]$ as an *arm*. We shall see, however, that this specific semi-bandit problem is challenging. In particular, the agent pays a *non-linear* search cost at each round (with respect to the selected combinatorial action), that additionally depends on the ordering. Moreover, due to the budget constraint, it is also an instance of *budgeted bandits*, also known as *bandits with knapsacks* (Badanidiyuru, Kleinberg, and Slivkins, 2013), in the case of single resource and infinite horizon. We thus evaluate the performance of a learning policy with the (common) notion of *budgeted regret*. It measures the expected difference, in terms of cumulative reward collected within the budget constraint B, between the learning policy and an *oracle* policy that knows a priori the exact parameters of the problem. Budgeted combinatorial semi-bandits have been already studied by

[2] One can notice that extending this work to the general case can be possible considering an α-approximation regret.

Sankararaman and Slivkins (2017) for several resources, but with a finite horizon. Moreover, their algorithm is efficient only for some specific combinatorial structures (such as matroids). The structure of constraints in sequential search-and-stop is in general more complex.

Motivation There are several motivations behind this setting. One example is the *decision-theoretic troubleshooting* problem of giving a diagnosis for several devices having a malfunctioning component and arriving sequentially to the agent. In many *troubleshooting* applications, we additionally face precedence constraints. These restrictions are imposed to the agent as the ordering of component tests, see e.g., Jensen et al., 2001. Moreover, allowing the agent to *stop* gives a new alternative to the so-called *service call* (Heckerman, Breese, and Rommelse, 1995; Jensen et al., 2001) in order to deal with non-cost-effective vertices: Instead of giving a high cost to an extra action that will automatically find the fault in the device, we give it a zero cost, but do not reward such diagnostic failure. This way, we do not need to estimate any *call-service* cost. This alternative is used, for example, when a new device is sent to the user if the diagnostic fails, with a cost that depends on a disutility for the user: loss of personal data, device reconfiguration, etc. Maximizing the number of hiders found is then analogous to maximizing the number of successful diagnoses.

Another example comes from online advertisement. There are several different actions that might generate a conversion from a user, such as sending one or several emails, displaying one or several ads on a website, buying keywords on search engines, etc. We assume that some precedence constraints are imposed between actions and that a conversion will occur if some sequence of actions is made, for instance, first, display an ad, then send the first email, and finally the second one. As a consequence, the conversion is "hidden", the precedence constraints restrict our access to it, and the agent aims at finding it. However, for some users, finding the correct sequence might be too expensive and it might be more interesting to abandon that specific user to focus on more promising ones.

Related settings Finally, there are several settings related to ours. One of them is stochastic probing (Gupta and Nagarajan, 2013), which differs in the fact that each arm can contain a hider, independently from each other. Another one is the machine learning framework of optimal discovery (Bubeck, Ernst, and Garivier, 2013).

Our contributions One of our main contributions is a stationary *offline policy* (i.e., an algorithm that solves the problem when the distribution is known), for which we prove the approximation guarantees and adapt it in order to fit the online problem. In particular, we prove that it is quasi-optimal and use SCHEDULING to prove its computational efficiency. Next, we provide a solution when the distribution is unknown to the agent, based on the *combinatorial upper confidence bounds* (CUCB) algorithm from Chen, Wang, and Yuan (2016), and UCB-variance (UCB-V) of Audibert, Munos, and Szepesvári (2009a). Dealing with variance estimates allows us to sharp the bound on the expected regret, improving the overall dependence on the dimension n compared to the simple use of CUCB. We also propose a new method (that can be of independent interest) to avoid the typical $c_{\min}^{-2}$ term in the expected regret bound (Tran-Thanh, Chapman, et al., 2012; Ding, Qin, et al., 2013; Xia, Ding, et al., 2016; Xia, Qin, et al., 2016; Watanabe et al., 2017), where $c_{\min}$ is the minimal expected search cost paid over a single round.

4.1.1 Background

We formalize in this subsection the setting we consider. We denote a finite DAG by $G \triangleq ([n], E)$, where $[n]$ is its set of vertices, or arms, and E is its set of directed edges. For more generality, we assume arm costs are random and mutually independent. We denote $C_j \in [0,1]$, with expectation $c_j^* \triangleq \mathbb{E}[C_j] > 0$, the cost of arm j. We thus have $\mathbf{c}^* = \mathbb{E}[\mathbf{C}] \in (0,1]^n$. We also assume that one specific vertex, called *hider*, is chosen at random, independently from $\mathbf{C}$, accordingly to some fixed categorical distribution (or Multinoulli) parameterized by vector $\mathbf{w}^*$ satisfying[3] $\sum_{i=1}^n w_i^* = 1$ and $w_i^* \in [0,1]$. Notice that $\mathbf{W} \sim \text{Multinoulli}(\mathbf{w}^*)$ if, given $i \in [n]$ and with probability w_i^*, $W_i = 1$ and $W_j = 0$ for all $j \neq i$.

Let $G\langle A\rangle$ be the sub-DAG in G induced by A, i.e., the DAG with A as vertex set, and with (i,j) an edge in $G\langle A\rangle$ if and only if $(i,j) \in E$. We call *support* of an ordered arm set $S = (s_1, \ldots, s_k)$ the corresponding non-ordered set. For two disjoint ordered arm sets S and S', we let $SS' = (s_1, s_2, \ldots, s_{|S|}, s'_1, s'_2, \ldots, s'_{|S'|})$ be the concatenation of S and S'.

We assume that G allows a polynomial-time algorithm (w.r.t. n), that takes some parameters $\mathbf{w}, \mathbf{c} \in \mathbb{R}_+^n$, and outputs $S = \textsc{scheduling}(\mathbf{w}, \mathbf{c}, G)$ minimizing

$$d(S; \mathbf{w}, \mathbf{c}) \triangleq \sum_{i=1}^{|S|} w_{s_i} \sum_{j=1}^{i} c_{s_j}$$

over linear extensions[4] $S = (s_1, \ldots, s_n)$ of the poset defined by G (that we call G-linear extensions). Notice that $d(S; \mathbf{w}^*, \mathbf{c}^*)$ represents the expected cost to pay for finding the hider with the G-linear extension S, i.e., by searching arm s_1 first and paying C_{s_1}, then s_2 by paying C_{s_2} in case $W_{s_1} = 0$, and so on until the hider is found, i.e., the last arm i searched is such that $W_i = 1$.

We define a *search* in G as an ordering $S = (s_1, \ldots, s_k)$ of different arms such that for all $i \in [k]$, predecessors of s_i in G are included in $\{s_1, \ldots, s_{i-1}\}$, i.e., a search is a prefix of a G-linear extension. We denote by $\mathcal{S}$ the set of searches in G. Search supports are called *initial sets*.

4.1.2 Protocol

Our search setting is *sequential*. We consider an *agent*, also called a *learning algorithm* or a *policy* that knows G but that does not know $\mathbb{P}_{(\mathbf{C},\mathbf{W})}$. At each round t, an independent sample $(\mathbf{C}_t, \mathbf{W}_t)$ is drawn from $\mathbb{P}_{(\mathbf{C},\mathbf{W})}$. The aim of the agent is to search the hider (i.e., the arm i such that $W_{i,t} = 1$) by constructing a *search* on G. Since the hider may be located at some arm that does not belong to the search, it is not necessarily found over each round.

The search to be used by the agent can be chosen based on all its previous observations, i.e., all the costs of explored vertices (and only those) and all the locations where the hider has been found or not. Obviously, the search cannot use the non-observed quantities. For example, the agent may estimate $\mathbf{w}^*$ and $\mathbf{c}^*$ in order to choose the search accordingly. Each time an arm j is searched, the feedback $W_{j,t}$ and $C_{j,t}$ is given to the agent. The agent can keep searching round after round until its global budget, B, runs out. B is a positive number and does not need to be known

[3] i.e., $\mathbf{w}^*$ belongs to the simplex of $\mathbb{R}^n$

[4] A linear extension of a poset is a total ordering consistent with the poset, i.e., if a is before b in the poset, then the same has to be true for its linear extension.

to the agent in advance. The agent wants to maximize the overall number of hiders found under the budget constraint.

The setting described above allows the agent to modify its behavior depending on the feedback it received during the current round. However, the only feedback on W_i susceptible to modify the search the agent chose at the beginning of a round t is the observation of $W_{i,t} = 1$ for some arm i. Even if nothing prevents the agent from continuing "searching" some arms after having seen such an event, it would not increase the number of hiders found (there is no more hider to find), while this would still decrease the remaining budget, and therefore it would have a pure exploratory purpose. Knowing this, for the case where costs are deterministic, an oracle policy that knows exactly $\mathbb{P}_{\mathbf{W}}$ thus *selects* a search S_t at the beginning of round t, and then *performs* the search that follows S_t until either $W_{i,t} = 1$ is observed or S_t is exhausted (i.e., no arms are left in S_t). Therefore, the performed search is composed of the set of arms

$$A_t = \{s_{t,1}, \dots, s_{t,j}\},$$

where j is such that $W_{s_j,t} = 1$, and $j = |S_t|$ if no such node exists. In the general case of random costs, when we will design a policy, we choose to restrict ourselves to agents that select a search S_t at the beginning of each round t and then performs A_t over this round. As a consequence, the selected search is computed based on observations collected during previous rounds $t-1, t-2, \dots$, denoted $\mathcal{H}_t$, that we refer to as *history*.

Remark 10. *Since a single arm i is associated to a pair of outcomes (C_i, W_i), it must be kept in mind that the above set A_t is the feedback set corresponding to the cost $\mathbf{C}$ only (it is on this feedback that the above setting is of the CMAB-T type). Concerning the feedback for $\mathbf{W}$, we remark that $\mathbf{e}_{S_t} \odot \mathbf{W}_t$ is observed at round t in the above setting, i.e., the feedback set coincides with the action selected (i.e., this feedback is only of the CMAB type).*

Following Stone (1976), we refer to our problem as *sequential search-and-stop*. We now detail the overall objective for this problem: The agent wants to follow a policy π, that performs a search A_t at round t, while maximizing the expected overall reward

$$F_B(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \sum_{i \in A_t} W_{i,t}\right],$$

where τ_B is the random round at which the remaining budget becomes negative: In particular, we have that if $B_t \triangleq B - \sum_{t'=1}^{t} \mathbf{e}_{A_{t'}}^{\intercal} \mathbf{C}_{t'}$, then $B_{\tau_B - 1} \geq 0$ and $B_{\tau_B} < 0$. We evaluate the performance of a policy using the *expected (budgeted) regret* with respect to F_B^*, the maximum value of F_B (among all possible oracle policies that know $\mathbb{P}_{(\mathbf{C},\mathbf{W})}$ and B), defined as

$$R_B(\pi) \triangleq F_B^* - F_B(\pi).$$

Example 3. *One may wonder if there exist cases where it is interesting for the agent to stop the search earlier (i.e. to select S with $|S| < n$). Consider for instance the simplest non-trivial case with two arms and no precedence constraint. The costs are deterministically chosen to be ε and 1 and the location of the hider is chosen uniformly at random. An optimal search will always first sample the arm with $\varepsilon < 1$ cost. If it also samples the other one, then the hider will be found with an expected cost of $\varepsilon + 1/2$. However, if the agent always stops the search after the first arm, and*

reinitializes on a new instance by doing the same, the overall cost to find one hider is

$$\sum_{t=1}^{\infty}\left(\frac{1}{2}\right)^t t\varepsilon = 2\varepsilon < \varepsilon + \frac{1}{2}, \quad \text{for } \varepsilon < \frac{1}{2}.$$

Therefore, stopping searches, even if the location of the hider is known, may be better than always trying to find it.

4.2 Offline oracle

In this section, we provide an algorithm for sequential search-and-stop when parameters $\mathbf{w}^*$ and $\mathbf{c}^*$ are given to the agent. We show that a simple stationary policy (i.e., the same search S^* is selected at each round) can obtain almost the same expected overall reward as F_B^*. We will denote by Oracle an algorithm that takes $\mathbf{w}^*$, $\mathbf{c}^*$, and G as input and outputs S^*. This *offline* oracle will eventually be used by the agent for the *online* problem, i.e., when parameters are unknown. Indeed, at round t, the agent can approximate S^* by the output S_t of Oracle$(\mathbf{w}_t, \mathbf{c}_t, G)$, where $\mathbf{w}_t, \mathbf{c}_t$ can be any guesses/estimates of the true parameters. Importantly, depending on the estimation followed by the agent, $\mathbf{w}_t$ may not stay in the simplex anymore. We will thus build Oracle such that an "acceptable" output is given for any input $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$.

4.2.1 Objective design

A standard paradigm for designing a stationary approximation of the offline problem in budgeted multi-armed bandits is the following: S^* has to minimize the ratio between the expected cost paid and the expected reward gain, over a single round, selecting S^*. We thus define, for $S \in \mathcal{S}$,

$$\begin{aligned} J(S; \mathbf{w}^*, \mathbf{c}^*) &\triangleq \mathbb{E}[\mathbf{e}_A^\intercal \mathbf{C}]\mathbb{E}[\mathbf{e}_S^\intercal \mathbf{W}]^{-1} \\ &= \frac{d(S; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{e}_S^\intercal \mathbf{w}^*)\mathbf{e}_S^\intercal \mathbf{c}^*}{\mathbf{e}_S^\intercal \mathbf{w}^*} \\ &= \sum_{i=1}^{|S|} \frac{c_{s_i}^*\left(1 - \mathbf{e}_{\{s_1,\dots,s_{i-1}\}}^\intercal \mathbf{w}^*\right)}{\mathbf{e}_S^\intercal \mathbf{w}^*}. \end{aligned}$$

Notice that we allow J to be equal to $+\infty$ (when $\mathbf{e}_S^\intercal \mathbf{w}^* = 0$). We use the convention $J(\emptyset; \mathbf{w}^*, \mathbf{c}^*) = +\infty$, because there is no interest in choosing an empty search for a round. We define the optimal values of J on $\mathcal{S}$ as

$$J^* \triangleq \min_{S \in \mathcal{S}} J(S; \mathbf{w}^*, \mathbf{c}^*), \quad \mathcal{S}^* \triangleq \arg\min_{S \in \mathcal{S}} J(S; \mathbf{w}^*, \mathbf{c}^*).$$

We now provide guarantees for this stationary policy.

Proposition 10. *If π^* is the offline policy selecting $S^* \in \mathcal{S}^*$ at each round t, then*

$$\frac{B - n}{J^*} \leq F_B(\pi^*) \leq F_B^* \leq \frac{B + n}{J^*}.$$

Proof. The proof follows the one provided for Lemma 1 of Xia, Qin, et al., 2016. If we let $B_0 = B$, then for any offline policy π, if we denote by A_t, S_t the search selected, resp. performed by π at round t (we saw that an optimal policy is such that S_t does not depend on $\mathbf{W}_t$), and if we let $B_t = B - \sum_{t'=1}^{t} \mathbf{e}_{A_{t'}}^\intercal \mathbf{C}_{t'}$ be the remaining budget

at time t,

$$
\begin{aligned}
F_B(\pi) &= \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S_t} \mathbb{I}\{B_t \geq 0,\ W_{i,t} = 1\}\right] && (4.1)\\
&\leq \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S_t} \mathbb{I}\{B_{t-1} \geq 0,\ W_{i,t} = 1\}\right] && (4.2)\\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S_t} \mathbb{I}\{B_{t-1} \geq 0\} w_i^*\right] && (4.3)\\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0\} \mathbf{w}^{*\mathsf{T}} \mathbf{e}_{S_t}\right]\\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0\} \frac{d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{w}^{*\mathsf{T}} \mathbf{e}_{S_t}) \mathbf{c}^{*\mathsf{T}} \mathbf{e}_{S_t}}{J(S_t; \mathbf{w}^*, \mathbf{c}^*)}\right]\\
&\leq \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0\} \frac{d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{w}^{*\mathsf{T}} \mathbf{e}_{S_t}) \mathbf{c}^{*\mathsf{T}} \mathbf{e}_{S_t}}{J^*}\right]\\
&= \frac{1}{J^*} \mathbb{E}\left[\sum_{t=1}^{\tau_B} (d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{w}^{*\mathsf{T}} \mathbf{e}_{S_t}) \mathbf{c}^{*\mathsf{T}} \mathbf{e}_{S_t})\right] && (4.4)\\
&= \frac{1}{J^*} \mathbb{E}\left[\sum_{t=1}^{\tau_B} \mathbf{c}^{*\mathsf{T}} \mathbf{e}_{A_t}\right]\\
&= \frac{1}{J^*} \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \mathbf{C}_t^{\mathsf{T}} \mathbf{e}_{A_t} + \mathbf{C}_{\tau_B}^{\mathsf{T}} \mathbf{e}_{A_{\tau_B}}\right]\\
&\leq \frac{B + n}{J^*}, && (4.5)
\end{aligned}
$$

where (4.2) uses $B_t \geq 0 \Rightarrow B_{t-1} \geq 0$, (4.3) is obtained by conditioning on previously sampled arms, (4.4) uses the random round τ_B such that $B_{\tau_B - 1} \geq 0$ and $B_{\tau_B} < 0$, and (4.5) uses the definition of $B_{\tau_B - 1}$ and $C_{i,t} \leq 1$. Now, for the lower bound, we have that

$$
\begin{aligned}
F_B(\pi^*) &\geq \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S^*} \mathbb{I}\{B_{t-1} \geq n,\ W_{i,t} = 1\}\right] && (4.6)\\
&= \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{I}\{B_{t-1} \geq n\} \mathbf{e}_{S^*}^{\mathsf{T}} \mathbf{w}^*] && (4.7)\\
&= \frac{1}{J^*} \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbf{C}_t^{\mathsf{T}} \mathbf{e}_{A_t^*}\right] && (4.8)\\
&\geq \frac{B - n}{J^*}, && (4.9)
\end{aligned}
$$

where (4.6) uses $B_{t-1} \geq n \Rightarrow B_t \geq 0$, (4.7) uses the same derivation as previously, (4.8) uses τ, the random round such that $B_{\tau - 1} \geq n$ and $B_\tau < n$, and (4.9) is by definition of B_τ. In (4.8), $A_t^* = \left\{s_1^*, \ldots, s_j^*\right\}$, where j is such that $W_{s_j,t} = 1$, and $j = |S^*|$ if no such node exists. □

Intuitively, Proposition 10 states that the optimal overall expected reward that can be gained (i.e., the maximum expected number of hiders found) is approximately B/J^* (we assume that $B \gg n$). This is quite intuitive, since this quantity is actually

the ratio between the overall budget and the minimum expected cost paid to find a *single* hider. Indeed, one can consider the related problem of minimizing the overall expected cost paid, over several rounds, to find a single hider. It can be expressed as an infinite-time horizon Markov decision process (MDP) with action space $\mathcal{S}$ and two states: whether the hider is found (which is the terminal state) or not. The goal is to choose a strategy $S_1, S_2, \dots, S_t, \dots$, minimizing

$$\begin{aligned}\mathcal{J}(S_1, S_2, \dots) &\triangleq \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbf{e}_{A_t}^\intercal \mathbf{C}_t\right] \\ &= \sum_{t=1}^{\infty}\left(\mathbf{e}_{S_t}^\intercal \mathbf{w}^* \left(\sum_{u=1}^{t-1} \mathbf{e}_{S_u}^\intercal\right)\mathbf{c}^* + d(S_t; \mathbf{w}^*, \mathbf{c}^*)\right)\prod_{u=1}^{t-1}(1 - \mathbf{e}_{S_u}^\intercal \mathbf{w}^*),\end{aligned}$$

where the stopping time τ is the first round at which the hider is found. The Bellman equation is

$$\mathcal{J}(S_1, S_2, \dots) = d(S_1; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{e}_{S_1}^\intercal \mathbf{w}^*)(\mathbf{e}_{S_1}^\intercal \mathbf{c}^* + \mathcal{J}(S_2, \dots)),$$

from which we deduce there exists an optimal *stationary* strategy (Sutton and Barto, 1998) such that $S_t = S$ for all $t \in \mathbb{N}^*$. Therefore, we can minimize $\mathcal{J}(S, S, \dots) = J(S; \mathbf{w}^*, \mathbf{c}^*)$ that gives the optimal value of J^*.

As we already mentioned, Oracle aims at taking inputs $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$. The first straightforward way to do is to consider

$$J(S; \mathbf{w}, \mathbf{c}) \triangleq \sum_{i=1}^{|S|} \frac{c_{s_i}\left(1 - \mathbf{e}_{\{s_1,\dots,s_{i-1}\}}^\intercal \mathbf{w}\right)}{\mathbf{e}_S^\intercal \mathbf{w}}.$$

However, notice that with the definition above, $J(\,\cdot\,; \mathbf{w}, \mathbf{c})$ could output negative values (if $\mathbf{e}_{[n]}^\intercal \mathbf{w} > 1$), which is not desired, because the agent would then be enticed to search arms with a high cost. We thus need to design a non-negative extension of J to $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$. One way is to replace $\left(1 - \mathbf{e}_{\{s_1,\dots,s_{i-1}\}}^\intercal \mathbf{w}\right)$ by $\mathbf{e}_{\{s_1,\dots,s_{i-1}\}^c}^\intercal \mathbf{w}$, another is to consider $0 \vee J(S; \mathbf{w}, \mathbf{c})$. There is a significant advantage of considering the second way, even if it is less natural than the first one, which is that for $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$,

$$0 \vee J(S; \mathbf{w}, \mathbf{c}) \leq J(S; \mathbf{w}^*, \mathbf{c}^*),$$

if $\mathbf{w} \geq \mathbf{w}^*$ and $\mathbf{c} \leq \mathbf{c}^*$. This property[5] is known to be useful for analysis of many stochastic combinatorial semi-bandit algorithms (see e.g., Chen, Wang, and Yuan, 2016). Thus, we choose for Oracle the minimization of the surrogate $0 \vee J(\,\cdot\,; \mathbf{w}, \mathbf{c})$.

4.2.2 Algorithm and guarantees

We now provide Oracle in Algorithm 3 and claim in Theorem 17 that it minimizes $0 \vee J(\,\cdot\,; \mathbf{w}, \mathbf{c})$ over $\mathcal{S}$. Notice that Oracle needs to call the polynomial-time algorithm $\textsc{scheduling}(\mathbf{w}, \mathbf{c}, G)$, that minimizes the objective function $d(S; \mathbf{w}, \mathbf{c})$ over G-linear extensions S. Then, Algorithm 3 only computes the maximum value index of a list of size n that takes linear time. To give an intuition, S^* follows the ordering given by $\textsc{scheduling}(\mathbf{w}, \mathbf{c}, G)$, and stops at some point when it becomes more interesting to start a fresh new instance.

[5] Notice this is not exactly a monotonicity property, because we compare to a single point $(\mathbf{w}^*, \mathbf{c}^*)$.

Algorithm 3 Oracle

Input: $\mathbf{w}, \mathbf{c}$ and G.
$S = \{s_1, \dots, s_n\} \triangleq \text{SCHEDULING}(\mathbf{w}, \mathbf{c}, G)$.
$i^* \triangleq \arg\min_{i \in [n]} 0 \vee J(\{s_1, \dots, s_i\}; \mathbf{w}, \mathbf{c})$ (ties may be broken arbitrarily).
Output: the search $S^* \triangleq \{s_1, \dots, s_{i^*}\}$.

Theorem 17. *For every* $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$, *Algorithm 3 outputs a search minimizing* $0 \vee J(\,\cdot\,; \mathbf{w}, \mathbf{c})$ *over* $\mathcal{S}$.

The proof of Theorem 17 mixes known concepts of scheduling theory, such as Sidney decomposition (Sidney, 1975), with some new results for our objective function.

Until the end of this subsection, we might abbreviate $0 \vee J(\,\cdot\,; \mathbf{w}, \mathbf{c})$ into J^+, $J(\,\cdot\,; \mathbf{w}, \mathbf{c})$ into J and $d(\,\cdot\,; \mathbf{w}, \mathbf{c})$ into d, keeping in mind that our results will be valid for all $(\mathbf{w}, \mathbf{c}) \in (\mathbb{R}_+^n)^2$. To prove Theorem 17 we first define the concept of *density*, well know in scheduling and search theory.

Definition 13 (Density)**.** *The density is the function defined on* $A \in \mathcal{P}([n])$ *by* $\lambda(A) \triangleq \mathbf{e}_A^\intercal \mathbf{w} / \mathbf{e}_A^\intercal \mathbf{c}$, *and* $\lambda(\emptyset) = 0$.

Density of $A \subset [n]$ can be understood as the quality/price ratio of that set of arms: the quality is the overall probability of finding the hider in it, while the price is the total cost to fully explore it. Without precedence constraint, the so-called Smith's rule of ratio (Smith, 1956) gives that S minimizes d over linear orders (i.e., permutations of $[n]$) if and only if $\lambda(s_1) \geq \cdots \geq \lambda(s_n)$.[6] Sidney (1975) generalized this principle to any precedence constraint with the concept of Sidney decomposition. Recall that an initial set is the support of a search.

Definition 14 (Sidney decomposition)**.** *A Sidney decomposition* $(Z_1, Z_2, \dots, Z_k)$ *is an ordered partition of* $[n]$ *such that for all* $i \in [k]$, Z_i *is an initial set of maximum density in* $G\langle Z_i \sqcup \cdots \sqcup Z_k \rangle$.

Notice that the Sidney decomposition defines a poset on $[n]$, with the constraint that an element of Z_i must be processed before those of Z_j for $i < j$. Any G-linear extension that is also a linear extension of this poset is said to be *consistent* with the Sidney decomposition. The following theorem was proved by Sidney (1975):

Theorem 18 (Sidney, 1975)**.** *Every minimizer of* d *over* G*-linear extensions is consistent with some Sidney decomposition. Moreover, for every Sidney decomposition* $(Z_1, \dots, Z_k)$, *there is a minimizer of* d *over* G*-linear extensions that is consistent with* $(Z_1, \dots, Z_k)$.

Notice that Theorem 18 does not provide a full characterization of minimizers of d over G-linear extensions, but only a *necessary* condition. Nothing is stated about how to chose the ordering inside each Z_i's, and this highly depends on the structure of G (Lawler, 1978; Ambühl and Mastrolilli, 2009; Ambühl, Mastrolilli, et al., 2011). We are now ready to prove Theorem 17, thanks to Lemma 5.

Lemma 5. *For any Sidney decomposition* $(Z_1, \dots, Z_k)$, *there exists* $i \leq k$ *and a search with support* $Z_1 \sqcup \cdots \sqcup Z_i$ *that minimizes* J^+.

[6] One can see that $\sum_{\{i,j\} \in I(\sigma),\ i<j} c_{s_i} c_{s_j} (\lambda(s_i) - \lambda(s_j))$ is the variation of d when swapping a linear order S by a permutation σ, where $I(\sigma)$ the set of inversions in σ.

Proof of Theorem 17. We know from first statement of Theorem 18 that

$$S \triangleq \textsc{scheduling}(\mathbf{w}, \mathbf{c}, G)$$

given in Algorithm 3 is consistent with some Sidney decomposition $(Z_1, \dots, Z_k)$. Let $i \leq k$ and S_* minimizing J^+ of support $Z_1 \sqcup \cdots \sqcup Z_i$ given by Lemma 5. Consider the following decomposition of S: $S = S'S''$ with S' being the restriction of S to $Z_1 \sqcup \cdots \sqcup Z_i$ (and thus S'' is its restriction to $Z_{i+1} \sqcup \cdots \sqcup Z_k$). Let's prove that S' is also a minimizer of J^+ by showing $J^+(S') \leq J^+(S_*)$, thereby concluding the proof. By definition of S, we have $0 \leq d(S_*S'') - d(S'S'') = d(S_*) - d(S')$, so

$$\frac{d(S') + (1 - \mathbf{w}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}) \mathbf{c}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}}{\mathbf{w}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}} \leq \frac{d(S_*) + (1 - \mathbf{w}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}) \mathbf{c}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}}{\mathbf{w}^\intercal \mathbf{e}_{Z_1 \sqcup \cdots \sqcup Z_i}},$$

i.e., $J(S') \leq J(S_*)$, and because $x \mapsto 0 \vee x$ is non-deacreasing on $\mathbb{R}$, we have $J^+(S') \leq J^+(S_*)$. □

The proof of Lemma 5 also uses Sidney's Theorem 18, but this time the second statement. However, although it provides a crucial analysis, with fixed support, concerning the order to choose for minimizing d and therefore J^+, nothing is said about the support to choose. Thus, to prove Lemma 5, we also need the following Proposition 11, that gives the key support property satisfied by J^+.

Proposition 11 (Support property)**.** *If $SS', SS'S'' \in \mathcal{S}$ with $\lambda(S'') \geq \lambda(S')$, then*

$$J^+(SS') \geq J^+(S) \wedge J^+(SS'S''). \tag{4.10}$$

Proof. If $J(SS'S'') < 0$, then $J^+(SS'S'') = 0 \leq J^+(SS')$ and (4.10) is true. We thus assume $J(SS'S'') \geq 0$. Since $J(S'') \leq \frac{1}{\lambda(S'')}$,

$$\begin{aligned} 0 \leq J(SS'S'') &= \frac{J(SS')\mathbf{w}^\intercal \mathbf{e}_{SS'}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} + \frac{\mathbf{w}^\intercal \mathbf{e}_{S''} J(S'') - \mathbf{w}^\intercal \mathbf{e}_{SS'} \mathbf{c}^\intercal \mathbf{e}_{S''}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} \\ &\leq \frac{J(SS')\mathbf{w}^\intercal \mathbf{e}_{SS'}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} + \frac{\mathbf{w}^\intercal \mathbf{e}_{S''}(1 - \mathbf{w}^\intercal \mathbf{e}_{SS'})}{\lambda(S'')\mathbf{w}^\intercal \mathbf{e}_{SS'S''}}. \end{aligned} \tag{4.11}$$

If $1 - \mathbf{w}^\intercal \mathbf{e}_{SS'} \leq 0$, by (4.11), we have that

$$0 \leq J(SS'S'') \leq \frac{J(SS')\mathbf{w}^\intercal \mathbf{e}_{SS'}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} \leq J(SS'),$$

so $J^+(SS'S'') \leq J^+(SS')$ and (4.10) is true. Thus, we suppose that $1 - \mathbf{w}^\intercal \mathbf{e}_{SS'} \geq 0$. If $J(S) \leq J(SS')$, then $J^+(S) \leq J^+(SS')$ and (4.10) is true. Else,

$$J(SS') \geq \frac{1}{\mathbf{w}^\intercal \mathbf{e}_{S'}} (J(SS')\mathbf{w}^\intercal \mathbf{e}_{SS'} - J(S)\mathbf{w}^\intercal \mathbf{e}_S) \geq \frac{1 - \mathbf{w}^\intercal \mathbf{e}_{SS'}}{\lambda(S')}. \tag{4.12}$$

Thus, we have

$$\begin{aligned} &J(SS'S'') - J(SS') \\ &\leq \frac{J(SS')\mathbf{w}^\intercal \mathbf{e}_{SS'}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} + \frac{\mathbf{w}^\intercal \mathbf{e}_{S''}(1 - \mathbf{w}^\intercal \mathbf{e}_{SS'})}{\lambda(S'')\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} - J(SS') \qquad (4.11) \end{aligned}$$

$$= \frac{-\mathbf{w}^\intercal \mathbf{e}_{S''} J(SS')}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} + \frac{\mathbf{w}^\intercal \mathbf{e}_{S''}(1 - \mathbf{w}^\intercal \mathbf{e}_{SS'})}{\lambda(S'')\mathbf{w}^\intercal \mathbf{e}_{SS'S''}}$$
$$\leq \frac{\mathbf{w}^\intercal \mathbf{e}_{S''}}{\mathbf{w}^\intercal \mathbf{e}_{SS'S''}} \left(\frac{-(1 - \mathbf{w}^\intercal \mathbf{e}_{SS'})}{\lambda(S')} + \frac{1 - \mathbf{w}^\intercal \mathbf{e}_{SS'}}{\lambda(S'')} \right) \leq 0 \quad \text{(4.12) and } \lambda(S'') \geq \lambda(S').$$

So, $J^+(SS'S'') \leq J^+(SS')$ and (4.10) is true. □

Example 4. *Now, as a preview, we can actually derive easily the proof of Lemma 5 when there is no precedence constraints, the idea in the general case being very similar. Let $(Z_1, \dots, Z_k)$ be a Sidney decomposition. Then, if $z_{i,1}, \dots, z_{i,j_i}$ are arms of Z_i, we have*

$$\lambda(z_{1,1}) = \cdots = \lambda(z_{1,j_1}) \geq \cdots \geq \lambda(z_{k,1}) = \cdots = \lambda(z_{k,j_k}).$$

Let S_* be a maximum-size minimizer of J^+ of support A_*. Assume A_* is not of the form given by Lemma 5, and let x be the first, for the order

$$(z_{1,1}, \dots, z_{1,j_1}, \dots, z_{k,1}, \dots, z_{k,j_k}),$$

in some $Z_i \backslash A_*$ while $A_* \cap (Z_i \sqcup \cdots \sqcup Z_k) \neq \emptyset$. By Proposition 11, we keep the optimality by either adding x to S_* (which contradicts the maximality of $|S_*|$), or by removing the suffix defined on $A_* \cap (Z_i \sqcup \cdots \sqcup Z_k)$, giving a support satisfying conclusion of Lemma 5.

4.2.3 Proof of Lemma 5

Before proving Lemma 5, we state some preliminaries about initial sets of the DAG G.

Fact 1. *A is an initial set in G if and only if for all $a \in A$, the predecessors of a in G are also in A.*

Let us recall that $\mathcal{L} \subset \mathcal{P}([n])$ is a lattice if $A, A' \in \mathcal{L} \Rightarrow (A \cap A', A \cup A' \in \mathcal{L})$.

Proposition 12. *The set of initial sets in G is a lattice.*

Proof. Let A and A' be two initial sets in G. If $a \in A \cup A'$ (respectively $a \in A \cap A'$), then the predecessors of a are included in predecessors of A or (respectively and) the predecessors of A', i.e., in A or (respectively and) A', so in $A \cup A'$ (respectively $A \cap A'$). □

Even if we do not use the following proposition,[7] we provide it nonetheless, since it illustrates how to handle density λ.

Proposition 13. *The set of initial sets of maximum density in G is a lattice.*

Proof. We use the fact that for $a, b \geq 0$ and $a', b' > 0$, $\frac{a+b}{a'+b'} \leq \frac{a}{a'} \vee \frac{b}{b'}$, with equality if and only if $\frac{a}{a'} = \frac{b}{b'}$. Indeed, if A and A' are two initial sets of maximum density in G, then

$$\frac{\mathbf{w}^\intercal \mathbf{e}_A}{\mathbf{c}^\intercal \mathbf{e}_A} = \frac{\mathbf{w}^\intercal (\mathbf{e}_A + \mathbf{e}_{A'})}{\mathbf{c}^\intercal (\mathbf{e}_A + \mathbf{e}_{A'})} = \frac{\mathbf{w}^\intercal (\mathbf{e}_{A \cup A'} + \mathbf{e}_{A \cap A'})}{\mathbf{c}^\intercal (\mathbf{e}_{A \cup A'} + \mathbf{e}_{A \cap A'})} \leq \frac{\mathbf{w}^\intercal \mathbf{e}_{A \cup A'}}{\mathbf{c}^\intercal \mathbf{e}_{A \cup A'}} \vee \frac{\mathbf{w}^\intercal \mathbf{e}_{A \cap A'}}{\mathbf{c}^\intercal \mathbf{e}_{A \cap A'}}.$$

$A \cap A'$ and $A \cup A'$ are initial sets, so by maximality of density of A,

$$\frac{\mathbf{w}^\intercal \mathbf{e}_{A \cup A'}}{\mathbf{c}^\intercal \mathbf{e}_{A \cup A'}} \vee \frac{\mathbf{w}^\intercal \mathbf{e}_{A \cap A'}}{\mathbf{c}^\intercal \mathbf{e}_{A \cap A'}} \leq \frac{\mathbf{w}^\intercal \mathbf{e}_A}{\mathbf{c}^\intercal \mathbf{e}_A}.$$

[7] Theorem 18 does need this proposition.

Therefore, the equality holds, and it needs to be the case that

$$\frac{\mathbf{w}^\intercal \mathbf{e}_{A\cup A'}}{\mathbf{c}^\intercal \mathbf{e}_{A\cup A'}} = \frac{\mathbf{w}^\intercal \mathbf{e}_{A\cap A'}}{\mathbf{c}^\intercal \mathbf{e}_{A\cap A'}} = \frac{\mathbf{w}^\intercal \mathbf{e}_{A}}{\mathbf{c}^\intercal \mathbf{e}_{A}},$$

so both $A \cap A'$ and $A \cup A'$ have maximum density. □

We now provide a proof of Lemma 5.

Proof of Lemma 5. Let j be the largest integer such that there is a search minimizing J^+ of the form $S = S^1 \cdots S^j S'$ with S^i of support Z_i for all $i \in [j]$. Let $S = S^1 \cdots S^j S'$ be such search, with $|S|$ being the smallest possible (i.e., $|S'|$ being the smallest possible). Let A' be the support of S'. By contradiction, assume $A' \neq \emptyset$. By Theorem 18, there exists a minimizer of the form $S^{j+1}S''$ of $d(S^1 \cdots S^j \cdot)$ over $G\langle Z_{j+1} \cup A'\rangle$-linear extensions, with S^{j+1} of support Z_{j+1}. $Z_{j+1} \cap A'$ is an initial set (as an intersection of two initial sets) of $G\langle Z_{j+1} \sqcup \cdots \sqcup Z_k\rangle$. Therefore, by maximiality of the density of Z_{j+1} in $G\langle Z_{j+1} \sqcup \cdots \sqcup Z_k\rangle$, we have

$$\lambda(Z_{j+1} \cap A') \leq \lambda(Z_{j+1}).$$

This translates, by property of the density, into

$$\lambda(Z_{j+1}) = \lambda((Z_{j+1} \cap A') \sqcup (Z_{j+1} \backslash A')) \leq \lambda(Z_{j+1} \backslash A'),$$

and thus $\lambda(A') \leq \lambda(Z_{j+1}) \leq \lambda(Z_{j+1} \backslash A')$. If we let S''' be a search of $G\langle (Z_{j+1} \backslash A') \sqcup Z_{j+2} \sqcup \cdots \sqcup Z_k\rangle$ with support $Z_{j+1} \backslash A'$ (one can take the order induced by Z_{j+1} by removing the arms of A'), then by Proposition 11, associated with $d(S^1 \cdots S^j S^{j+1} S'') \leq d(S^1 \cdots S^j S' S''')$, we have that

$$J^+(S) \geq J^+(S^1 \cdots S^j) \wedge J^+(S^1 \cdots S^j S' S''') \geq J^+(S^1 \cdots S^j) \wedge J^+(S^1 \cdots S^j S^{j+1} S''),$$

contradicting either the definition of j or the minimality of $|S|$. □

4.3 Online search-and-stop

In this section, we consider an additional challenge where the distribution $\mathbb{P}_{(\mathbf{C},\mathbf{W})}$ is unknown and the agent must deal with it, while minimizing $R_B(\pi)$ over sampling policies π, where B is a fixed budget. Recall that a policy π selects a search $S_t = \left\{s_{1,t}, \ldots, s_{|S_t|,t}\right\}$ at the beginning of round t, using previous observations $\mathcal{H}_t$, and then performs the search that choose the set of arms $A_t = \{s_{1,t}, \ldots, s_{j,t}\}$, where $W_{j,t} = 1$ or $j = |S_t|$. We treat the setting as a variant of stochastic combinatorial semi-bandits (Gai, Krishnamachari, and Jain, 2012). The feedback received by an agent at round t is random, because it depends on $\mathbf{W}_t$, and thus it is not measurable w.r.t. $\mathcal{H}_t$. More precisely, $(W_{i,t}, C_{i,t})$ is observed only for arms $i \in A_t$. Notice that since $\mathbf{W}_t$ is a one-hot vector, the agent can always deduce the value of $W_{i,t}$ for all $i \in S_t$. As a consequence, we will maintain two types of counters for all arms $i \in [n]$ and all $t \geq 1$,

$$\begin{aligned} N_{\oplus i,t-1} &\triangleq \sum_{t'=1}^{t-1} \mathbb{I}\{i \in S_{t'}\}, \\ N_{\ominus i,t-1} &\triangleq \sum_{t'=1}^{t-1} \mathbb{I}\{i \in A_{t'}\}. \end{aligned} \tag{4.13}$$

$$\begin{aligned} \overline{w}_{i,t-1} &\triangleq \frac{\sum_{t'=1}^{t-1} \mathbb{I}\{i \in S_{t'}\} W_{i,t'}}{N_{\oplus i,t-1}}, \\ \overline{c}_{i,t-1} &\triangleq \frac{\sum_{t'=1}^{t-1} \mathbb{I}\{i \in A_{t'}\} C_{i,t'}}{N_{\ominus i,t-1}}. \end{aligned} \tag{4.14}$$

We propose an approach similar to UCB-V of Audibert, Munos, and Szepesvári (2009a), based on CUCB of Chen, Wang, and Yuan (2016), called CUCB-V, that uses a variance estimation of $\mathbf{w}^*$ in addition to the empirical average. Notice that the variance of W_i for an arm i is $\sigma_i^2 \triangleq w_i^*(1 - w_i^*)$. Furthermore, since W_i is binary, the empirical variance of W_i after t rounds is $\overline{w}_{i,t}(1 - \overline{w}_{i,t})$. For every round t and every edge $i \in [n]$, with the previously defined empirical averages, we use the confidence bounds[8] defined as

$$\begin{aligned} c_{i,t} &\triangleq 0 \vee \left(\overline{c}_{i,t-1} - \sqrt{\frac{0.5\zeta \log t}{N_{\ominus i,t-1}}} \right), \\ w_{i,t} &\triangleq \left(\overline{w}_{i,t-1} + \sqrt{\frac{2\zeta \overline{w}_{i,t-1}(1 - \overline{w}_{i,t-1}) \log t}{N_{\oplus i,t-1}}} + \frac{3\zeta \log t}{N_{\oplus i,t-1}} \right) \wedge 1, \end{aligned}$$

where we choose the exploration factor to be $\zeta \triangleq 1.2$. Notice that we could take any $\zeta > 1$ as shown by Audibert, Munos, and Szepesvári (2009a). We provide the policy $\pi_{\text{CUCB-V}}$ that we consider in Algorithm 4.

Algorithm 4 Combinatorial upper confidence bounds with variance estimates (CUCB-V) for sequential search-and-stop

Input: G.
for $t = 1..\infty$ **do**
  Select S_t given by Oracle$(\mathbf{w}_t, \mathbf{c}_t, G)$.
  Perform the search that follows S_t until the hider is found, i.e., sample the arms of A_t.
  Collect feedback and update counters and empirical averages according to (4.13) and (4.14).
end for

4.3.1 Analysis

Notice that since an arm $i \in S_t$ is pulled (and thus $C_{i,t}$ is revealed to the agent) with probability $1 - \mathbf{e}_{\{s_{1,t},\dots,s_{i-1,t}\}}^{\mathsf{T}} \mathbf{w}^*$ over round t, we fall, as anticipated in Remark 10, into the setting of *probabilistically triggered arms* w.r.t. costs, described by Chen, Wang, and Yuan (2016) and Wang and Chen (2017). Thus we could rely on these prior results. However, the main difficulty in our setting is that we also need to deal with probabilities $W_{i,t}$, that the agent actually observes for every arm i in S_t, either because it actually pulls arm i, or because it deduces the value (that is thus 0) from other pulls of round t. In particular, if we follow the analysis of Chen, Wang, and Yuan (2016) and Wang and Chen (2017), the double sum in the definition of J leads to expected regret bound that is quite large. Indeed, assuming that all costs are deterministically equal to 1, if we suffer an error of δ when approximating each w_i^*, then the global error can be as large as $\sum_{i=1}^{n} \sum_{j=1}^{i-1} \delta = \mathcal{O}(n^2\delta)$, contrary to just

[8] With the convention $x/0 = +\infty,\ \forall x \geq 0$.

$\mathcal{O}(n\delta)$ for the approximation error w.r.t. costs, that is more common in combinatorial semi-bandits. Thus, we rather combine their work with the variance estimates of w_i^*. Often, this does not provide a significant improvement over UCB in terms of expected regret (otherwise we could do the same for the costs), but since in our case, the variance is of order $1/n$, the gain is non-negligible.[9] We let $c_{\min} > 0$ be any deterministic lower bound on the set $\left\{\mathbf{e}_{A_t}^\intercal \mathbf{c}^*, t \geq 1\right\}$. Furthermore, we let

$$T_B \triangleq \lceil 2B/c_{\min} \rceil$$

and for any search S, we define the *gap* of S as

$$\begin{aligned}\Delta(S) &\triangleq \mathbf{e}_S^\intercal \mathbf{w}^* \left(\frac{J(S; \mathbf{w}^*, \mathbf{c}^*)}{J^*} - 1 \right) \\ &= \frac{1}{J^*} \sum_{i=1}^{|S|} c_{s_i}^* \left(1 - \sum_{j=1}^{i-1} w_{s_j}^* \right) - \sum_{i=1}^{|S|} w_{s_i}^* \geq 0,\end{aligned}$$

that represents the *local regret* of selecting a sub-optimal search S at some round. In addition, for each arm $i \in [n]$, we define

$$\Delta_{i,\min} \triangleq \inf_{S \notin \mathcal{S}^*:\ i \in S} \Delta(S) > 0.$$

We provide bounds for the expected regret of $\pi_{\text{CUCB-V}}$ in Theorem 19. The first bound is gap-dependent, and is characterized by $c_{\min}$, J^*, and σ_i^2 , $\Delta_{i,\min}$, $i \in [n]$. Its main term scales logarithmically w.r.t. B. The second bound is true for any sequential search-and-stop problem instance having a fixed value of $c_{\min} > 0$ and $J^* > 0$.

Theorem 19. *The expected regret of* CUCB-V *satisfies*

$$R_B(\pi_{\text{CUCB-V}}) = \mathcal{O}\left(n \log T_B \sum_{i \in [n]} \frac{1 + (J^* + n)^2 \sigma_i^2}{J^{*2} \Delta_{i,\min}} + \frac{(J^* + n)}{J^* n} \log\left(\frac{n \Delta_{i,\max}}{\Delta_{i,\min}} \right) \right).$$

In addition,

$$\sup R_B(\pi_{\text{CUCB-V}}) = \mathcal{O}\left(\sqrt{n} \left(1 + \frac{n}{J^*} \right) \sqrt{T_B \log T_B} \right),$$

where the sup *is taken over all possible sequential search-and-stop problems with fixed* $c_{\min}$ *and* J^*.

In the proof (that can be found in subsection 4.5.1), the main challenge comes from the estimation of $\mathbf{w}^*$ and not from $\mathbf{c}^*$. We recall that our analysis does not use *triggering probability groups* of Wang and Chen, 2017 for dealing with costs, since we succeed in dealing with the triggering probability by seeing it as the expectation of the event where the corresponding counter is updated (see the end of Chapter 3). For hider probabilities, there is no triggering probability (they are equal to 1).[10] Notice, the analysis of Wang and Chen, 2017 only considers a deterministic horizon. In our case, we need to deal with a *random-time* horizon. For that, notice that the regret upper bounds that hold in expectation are obtained by splitting the expectation into

[9] The error δ is thus scaled by the standard deviation, of order $1/\sqrt{n}$, giving a global error of $\mathcal{O}(n^{1.5}\delta)$. We therefore recover the factor $n^{1.5}$ given in Theorem 19.

[10] When we select search S, all feedback W_i, $i \in S$ is received with probability 1, so *triggering probabilities* are not useful.

two parts. The first part is filtered with a high-probability event on which the regret grows as the logarithm of the random horizon and the second one is filtered with a low-probability event, on which we bound the regret by a constant. Since the log function is concave, we can upper bound the expected regret by a term growing as the logarithm of the expectation of the random horizon, with Jensen's inequality. Finally, we upper bound the expectation of the random horizon to get the rate of $\log T_B$.

4.3.2 Tightness of our regret bounds

Since we succeeded in reducing the dependence on n in the expected regret with confidence bounds based on variance estimates, we can now ask whether this dependence in Theorem 19 is tight. We stress that our solution to sequential search-and-stop is *computationally efficient*. In particular, *both* the offline oracle optimization and the computation of the optimistic search S_t in the online part *are tractable*.

Whenever rewards are not arbitrary correlated (as is the case in our setting), we can potentially exploit these correlations in order to reduce the gap-dependent regret's dependence on n even further. This could be done by choosing a tighter confidence region such as a confidence ellipsoid (Degenne and Perchet, 2016b), or a KL-confidence ball (Combes et al., 2015) instead of coordinate-wise confidence intervals. Unfortunately, these do not lead to computationally efficient algorithms. It also seems from our Theorem 20 that there is an extra $\sqrt{n}$ factor in our gap-free bound. It is an open question whether a better gap-free regret bound exists.

To show that we are only a $\sqrt{n}$ factor away, in the following theorem we provide a class of sequential search-and-stop problems (parameterized by n and B) on which the regret bound provided in Theorem 19 is tight up to a $\sqrt{n}$ factor (and a logarithmic one).

Theorem 20. *For simplicity, let us assume that n is even and that B is a multiple of n. For any optimal online policy π, there is a sequential search-and-stop problem with n arms and budget B such that*

$$-4 + \frac{1}{28}\sqrt{\frac{B}{n}} \leq R_B(\pi) = \mathcal{O}\left(\sqrt{B \log\left(\frac{B}{n}\right)}\right).$$

For the proof (given in subsection 4.5.2), we consider a DAG composed of two disjoint paths (Figure 4.1), with all costs deterministically set to 1 and with the hider located either at $s_{n/2}$ or $s'_{n/2}$. This information is given to the agent. We then reduced this setting to a two-arm bandit over at least B/n rounds.

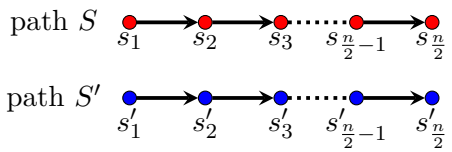

FIGURE 4.1: The DAG considered in Theorem 20.

Notice that bounds provided in Theorem 20 *decrease* with n. This is because, in the sequential search-and-stop problem, the increasing dependence on n is counterbalanced by the fact that the number of rounds is of order B/n, and that J^* is of order n.

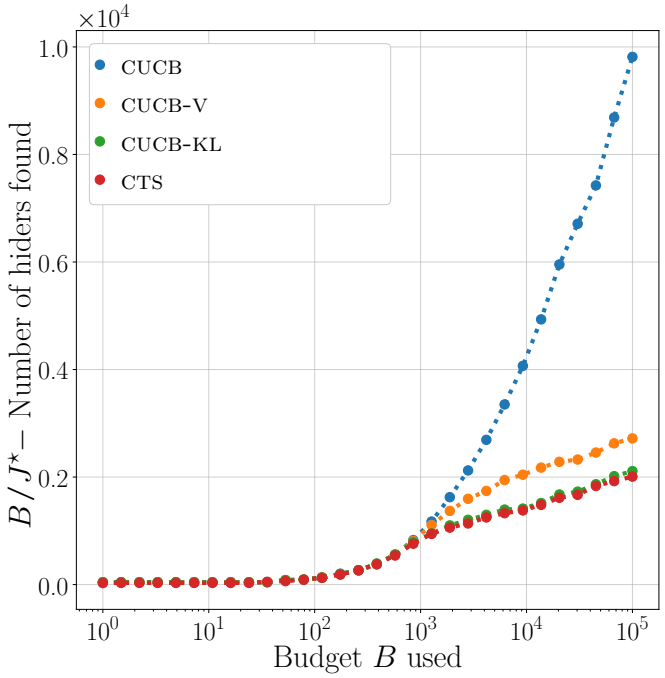

FIGURE 4.2: Cumulative regret for sequential search-and-stop, with B up to 10^5, averaged over 100 independent simulations.

Algorithm	Definition of $w_{i,t}$
CUCB	$\left(\overline{w}_{i,t-1} + \sqrt{\frac{0.5\zeta \log t}{N_{\oplus i,t-1}}}\right) \wedge 1$
CUCB-KL	The unique solution x to $N_{\oplus i,t-1}\mathrm{kl}(\overline{w}_{i,t-1}, x) = \zeta \log t$ such that $x \in [\overline{w}_{i,t-1}, 1]$
CTS	An independent sample from $\mathrm{Beta}(\alpha, N_{\oplus i,t-1} - \alpha)$, where $\alpha = N_{\oplus i,t-1}\overline{w}_{i,t-1}$

TABLE 4.1: Comparison algorithms in the experiment.

4.4 Experiments and discussion

In this section, we present an experiment for sequential search-and-stop. We compare our CUCB-V with three other online algorithms, which are same as CUCB-V except for the estimator $\mathbf{w}_t$ to be plugged in Oracle. We give corresponding definitions of $\mathbf{w}_t$ in Table 4.1, where we take $\zeta \triangleq 1.2$, and where

$$\mathrm{kl}(p, q) \triangleq p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

is the Kullback-Leibler divergence between two Bernoulli distributions of parameters $p, q \in [0, 1]$ respectively. In this table, CTS stands for *combinatorial Thompson sampling* (Wang and Chen, 2018).

We run simulations for all the algorithms with $n = 100$ and without precedence constraints, i.e., when the DAG is an edgeless graph. Notice that in this case, a search can be any ordered subset of arms (thus, the set of possible searches is of

cardinality $\sum_{k=0}^{n} n!/k! \leq en!$). This restriction does not remove complexity from the online problem, but rather from the offline one, so even in that case, the online problem is challenging. We take parameter $\mathbf{w}^*$ defined as

$$\begin{cases} w_i^* = \dfrac{1}{2^i} & \text{for } i \in [m-1] \\ w_m^* = \left(\dfrac{1}{2} + \varepsilon\right) w_{m-1}^* & \\ w_i^* = \left(\dfrac{1}{2} - \varepsilon\right) \dfrac{w_{m-1}^*}{n-m} & \text{for } i \in \{m+1, \dots, n\}, \end{cases}$$

where we chose $m \triangleq 40$. For $\varepsilon \in (0, 1/2)$, and for all expected cost being equal to each other, one can see that $\mathcal{S}^* = \{[m]\}$. Intuitively, w_i^* models the proportion of users answering i to some fixed request:[11] When $\varepsilon = 0$, half of the population answers 1, a quarter answers 2, ..., until m, and remaining users answer uniformly on remaining arms $\{m+1, \dots, n\}$. We chose $\varepsilon = 0.1$, $c_i^* = 1/2$ for all $i \in [n]$ and take $C_i \sim \text{Bernoulli}(c_i^*)$. In Figure 4.2, for each algorithm considered, we plot the quantity

$$\frac{B}{J^*} - \sum_{t=1}^{\tau_B - 1} \mathbf{e}_{S_t[\mathbf{W}_t]}^\intercal \mathbf{W}_t,$$

with respect to budget B, averaged over 100 simulations. As shown in Proposition 10, the curves obtained this way provide good approximations to the true regret curves. We notice that CUCB-KL, CUCB-V, and CTS are significantly better than CUCB, since the latter explores too much. In addition, the regret curves of CUCB-KL, CUCB-V and CTS are quite similar. In particular, their asymptotic slopes seem equal, which hints that regret rates are comparable on this instance.

Finally, let us note that it is not surprising that CUCB-KL outperforms CUCB-V, as we have binary variables here (cf. Chapter 2). More precisely, a CUCB-KL analysis would have given a slight improvement in the regret (using the kl function), but that the dependency in the dimension n does not change.

4.4.1 Discussion and future work

We presented sequential search-and-stop problem and provided a stationary offline solution. We gave theoretical guarantees on its optimality and proved that it is computationally efficient. We also considered the learning extension of the problem where the distribution of the hider and the cost are not known. We provided CUCB-V, an upper-confidence bound approach, tailored to our case and gave expected regret guarantees with respect to the optimal policy.

We now discuss several possible extensions of our work. We could consider several hiders rather than just one. Another would be to explore the CTS approach (Chapelle and Li, 2011; Agrawal and Goyal, 2012a; Komiyama, Honda, and Nakagawa, 2015; Wang and Chen, 2018) further in the learning case by considering a Dirichlet prior on the *whole* arm set. The Dirichlet seems appropriate because a sample $\mathbf{w}$ from this prior is in the simplex. The main drawback however is the difficulty of *efficiently* updating such prior to get the posterior, because in the case when the hider is not found, the one-hot vector is not received entirely.

[11] For recommender systems or search engines, w_i^* can thus be seen as the probability that a user aims to find i when entering the request.

4.5 Missing proofs

4.5.1 Proof of Theorem 19

Proof of Theorem 19. We start with showing a lower bound on the expected reward of any policy π,

$$
\begin{aligned}
F_B(\pi) &\geq \sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n\}\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*\big] && (4.15)\\
&= \sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n,\ S_t\in\mathcal{S}^*\}\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*\big] + \sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*\big]\\
&= \frac{1}{J^*}\sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n,\ S_t\in\mathcal{S}^*\}\big(d(S_t;\mathbf{w}^*,\mathbf{c}^*) + (1-\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*)\mathbf{e}_{S_t}^\mathsf{T}\mathbf{c}^*\big)\big]\\
&\quad + \frac{1}{J^*}\sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\big(d(S_t;\mathbf{w}^*,\mathbf{c}^*) + (1-\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*)\mathbf{e}_{S_t}^\mathsf{T}\mathbf{c}^*\big)\big]\\
&\quad - \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\Delta(S_t)]\\
&= \frac{1}{J^*}\sum_{t\geq 1} \mathbb{E}\big[\mathbb{I}\{B_{t-1}\geq n\}\big(d(S_t;\mathbf{w}^*,\mathbf{c}^*) + (1-\mathbf{e}_{S_t}^\mathsf{T}\mathbf{w}^*)\mathbf{e}_{S_t}^\mathsf{T}\mathbf{c}^*\big)\big]\\
&\quad - \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\Delta(S_t)]\\
&\geq \frac{B-n}{J^*} - \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\Delta(S_t)], && (4.16)
\end{aligned}
$$

with (4.15) obtained as (4.6) and (4.7), and (4.16) as (4.9). Therefore, since $F_B^*\leq (B+n)/J^*$ by Proposition 10, we have that

$$
\begin{aligned}
R_B(\pi) - \frac{2n}{J^*} &\leq \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1}\geq n,\ S_t\notin\mathcal{S}^*\}\Delta(S_t)]\\
&\leq \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_t\geq 0\}\Delta(S_t)] = \mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\Delta(S_t)\right].
\end{aligned}
$$

It is thus sufficient to bound this last expectation, which has the form of the classical regret (the non-budgeted one), and is thus easier to handle (note that we find here a random horizon, as anticipated in the end of the previous chapter).

For some round $t\geq 1$, recall that S_t minimizes $0\vee J(\,\cdot\,;\mathbf{w}_t,\mathbf{c}_t)$. In particular,

$$
0\vee J(S_t;\mathbf{w}_t,\mathbf{c}_t)\leq 0\vee J(S^*;\mathbf{w}_t,\mathbf{c}_t).
$$

We define the event

$$
\mathfrak{M}_t \triangleq \{\mathbf{w}_t\geq\mathbf{w}^*,\mathbf{c}_t\leq\mathbf{c}^*\},
$$

under which $0\vee J(S^*;\mathbf{w}_t,\mathbf{c}_t)\leq J^*$: indeed, we can first use $\mathbf{c}_t\leq\mathbf{c}^*$ to write

$$
J(S^*;\mathbf{w}^*,\mathbf{c}_t)\leq J(S^*;\mathbf{w}^*,\mathbf{c}^*) = J^*
$$

because $\mathbf{w}^*$ belongs to the simplex (thus, $1-\mathbf{e}_{\{s_1,\dots,s_{i-1}\}}^\mathsf{T}\mathbf{w}^*\geq 0$ for all i). Then, using $\mathbf{c}_t\geq 0$ with $\mathbf{w}_t\geq\mathbf{w}^*$, we can write $J(S^*;\mathbf{w}_t,\mathbf{c}_t)\leq J(S^*;\mathbf{w}^*,\mathbf{c}_t)$. The result follows since $x\mapsto x\vee 0$ is non-deacreasing on $\mathbb{R}$. As a consequence, under the event $\mathfrak{M}_t$, we

have

$$0 \vee J(S_t; \mathbf{w}_t, \mathbf{c}_t) \leq J^*.$$

We thus have that $\Delta(S_t)$ equals

$$\begin{aligned}
&\frac{1}{J^*}\left(\sum_{i=1}^{|S_t|} c^*_{s_{i,t}}\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) - J^*\mathbf{w}^{*\mathsf{T}}\mathbf{e}_{S_t}\right) \\
&= \frac{1}{J^*}\left(\sum_{i=1}^{|S_t|} c^*_{s_{i,t}}\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) - J^*\mathbf{w}_t^{\mathsf{T}}\mathbf{e}_{S_t}\right) + (\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t} \\
&\leq \frac{1}{J^*}\left(\sum_{i=1}^{|S_t|} c^*_{s_{i,t}}\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) - 0 \vee J(S_t; \mathbf{w}_t, \mathbf{c}_t)\mathbf{w}_t^{\mathsf{T}}\mathbf{e}_{S_t}\right) + (\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t} \\
&\leq \frac{1}{J^*}\left(\sum_{i=1}^{|S_t|} c^*_{s_{i,t}}\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) - J(S_t; \mathbf{w}_t, \mathbf{c}_t)\mathbf{w}_t^{\mathsf{T}}\mathbf{e}_{S_t}\right) + (\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t} \\
&= \frac{1}{J^*}\sum_{i=1}^{|S_t|}\Big(c^*_{s_{i,t}}\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) - c_{s_{i,t},t}\Big(1 - \mathbf{w}_t^{\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big)\Big) \\
&\quad + (\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t} \\
&= \frac{1}{J^*}\left(\sum_{i=1}^{|S_t|}\Big(c^*_{s_{i,t}} - c_{s_{i,t},t}\Big)\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) + \sum_{i=1}^{|S_t|} c_{s_{i,t},t}(\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\right) \\
&\quad + (\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t} \\
&\leq \frac{1}{J^*}\left(\sum_{i=1}^{|S_t|}\Big(c^*_{s_{i,t}} - c_{s_{i,t},t}\Big)\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big) + (n + J^*)(\mathbf{w}_t - \mathbf{w}^*)^{\mathsf{T}}\mathbf{e}_{S_t}\right).
\end{aligned}$$

We now define some more events:

$$\mathfrak{N}_{1,t} \triangleq \left\{\forall i \in S_t,\ c^*_i - c_{i,t} \leq 1 \wedge \sqrt{\frac{2\zeta \log t}{N_{\ominus i,t-1}}}\right\}.$$

$$\mathfrak{N}_{2,t} \triangleq \left\{\forall i \in S_t,\ w_{i,t} - w^*_i \leq 1 \wedge \left(\sqrt{\frac{8\zeta\sigma_i^2 \log t}{N_{\oplus i,t-1}}} + \frac{\frac{13}{3}\zeta \log t}{N_{\oplus i,t-1}}\right)\right\}.$$

$$\mathfrak{N}_t \triangleq \mathfrak{N}_{1,t} \wedge \mathfrak{N}_{2,t}$$

Under $\mathfrak{N}_t$, we further have

$$\begin{aligned}
\Delta(S_t) \leq \frac{1}{J^*}\sum_{i=1}^{|S_t|} \underbrace{\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big)}_{\mathbb{P}[\,s_i \in A_t|\mathcal{F}_t]} 1 \wedge \sqrt{\frac{2\zeta \log t}{N_{\ominus s_i,t-1}}} \\
+ \frac{n + J^*}{J^*}\sum_{i \in S_t} 1 \wedge \left(\sqrt{\frac{8\zeta\sigma_i^2 \log t}{N_{\oplus i,t-1}}} + \frac{\frac{13}{3}\zeta \log t}{N_{\oplus i,t-1}}\right).
\end{aligned}$$

On can notice that the factor $\Big(1 - \mathbf{w}^{*\mathsf{T}}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}\Big)$ is actually the probability of observing the cost of s_i, which is $\mathbb{P}[\,s_i \in A_t|\mathcal{F}_t] = \mathbb{P}[\,s_i \in A_t|\mathcal{H}_t]$ (since the policy is not randomized). If we summarize from that point on, we used the events $\mathfrak{M}_t$ and $\mathfrak{N}_t$ to prove the above upper bound on $\Delta(S_t)$. So we can break down the regret into

two pieces, whether or not $\mathfrak{M}_t \wedge \mathfrak{N}_t$ occurs.

Let's first treat the case where it occurs (first piece): We want to use Theorem 12 from Chapter 3. First of all, see that $\tau_B - 1$ is such that $\{t \leq \tau_B - 1\} \in \mathcal{F}_{t+1}$, so we use the horizon τ_B instead of $\tau_B - 1$ in this theorem. We can see in the above bound that we have a composed bonus (with 3 terms). In addition to that, we can take apart the null counters to have an initialization term of the form of the one in Proposition 5. We can thus use Proposition 6 to treat each of this four term separately, each one giving a final term in the regret. From Proposition 5, the initialization term induces the regret term $\sum_{i\in[n]} \frac{2}{J^*}$. Theorem 12 can be used for the three other terms (with $b_i(S_t)$ being a constant):

- with $\beta_{i,\tau_B} = \log(\tau_B)/J^{*2}$, $\alpha_i = 1/2$ and $p_{s_j}(S_t) = 1 - \mathbf{w}^{*\intercal}\mathbf{e}_{\{s_{t,1},\dots,s_{t,i-1}\}}$, the first term induces a regret term of order

$$\sum_{i\in[n]} \frac{n\mathbb{E}[\log(\tau_B)]}{J^{*2}\Delta_{i,\min}}.$$

- with $\beta_{i,\tau_B} = (n + J^*)^2\sigma_i^2 \log(\tau_B)/J^{*2}$, $\alpha_i = 1/2$ and $p_{s_j}(S_t) = 1$, the second term induces a regret term of order

$$\sum_{i\in[n]} \frac{n(n + J^*)^2\sigma_i^2\mathbb{E}[\log(\tau_B)]}{J^{*2}\Delta_{i,\min}}.$$

- with $\beta_{i,\tau_B} = (n + J^*)\log(\tau_B)/J^*$, $\alpha_i = 1$ and $p_{s_j}(S_t) = 1$, the last term induces a regret term of order

$$\sum_{i\in[n]} \frac{(n + J^*)\mathbb{E}[\log(\tau_B)]}{J^*} \log\left(\frac{n\Delta_{i,\max}}{\Delta_{i,\min}}\right).$$

We will now turn to the second piece. For this, we are going to use several Theorems from Chapter 2. In particular, we can use Corollary 1 (Hoeffding's inequality) and Theorem 8 (empirical Bernstein inequality) to get that $\mathfrak{M}_t$ holds with probability at least $1 - 4n\left\lceil\zeta\log^2(t)\right\rceil t^{-\zeta}$. On the other hand, using the other inequality from Corollary 1, Theorem 6 (Bernstein's inequality) and Proposition 3, we get that $\mathfrak{N}_t$ holds with probability at least $1 - 3n\left\lceil\zeta\log^2(t)\right\rceil t^{-\zeta}$. Therefore, the event $\mathfrak{M}_t \wedge \mathfrak{N}_t$ holds with probability at least $1 - 7n\left\lceil\zeta\log^2(t)\right\rceil t^{-\zeta}$ and the regret under the complementary event can be bounded by

$$7n\Delta_{\max}\sum_{t\geq 1}\left\lceil\zeta\log^2(t)\right\rceil t^{-\zeta} < \infty.$$

This terminates the analysis for the gap-dependent regret bound (we note that there's still $\mathbb{E}[\log(\tau_B)]$ to be bounded, which we'll do right after we deal with the gap-free regret bound).

Gap-free bound For the gap-free bound, we consider the event that the gap $\Delta(S_t)$ is greater than

$$\frac{(n + J^*)\sqrt{n\sum_{i\in[n]}\sigma_i^2\log(\tau_B)}}{J^*\tau_B^{1/2}}. \tag{4.17}$$

Under this event, we have the same analysis as above, but with $\Delta_{i,\min}$ replaced by the quantity (4.17). The obtained rate is the same under the complementary event, since the regret is thus trivially bounded by τ_B times the quantity (4.17). The result thus follows noticing that $\sum_{i\in[n]} \sigma_i^2 \leq 1$.

Control on the random time horizon We note that in both the gap-free and gap-dependent bound above, one must control the expectation of a concave increasing function of τ_B. Thus, with Jensen's inequality, we have to control $\mathbb{E}[\tau_B]$. This is the purpose of the following.

$$\begin{aligned}
\mathbb{E}[\tau_B] = 1 + \mathbb{E}\left[\sum_{t\geq 1} \mathbb{I}\{B_t \geq 0\}\right] &= 1 + \sum_{t\geq 1} \mathbb{P}\left[B - tc_{\min} + tc_{\min} \geq \sum_{t'=1}^{t} \mathbf{C}_{t'}^{\mathsf{T}} \mathbf{e}_{A_{t'}}\right] \\
&\leq T_B + 1 + \sum_{t\geq T_B+1} \exp\left(\frac{-2(B - tc_{\min})^2}{t}\right) \qquad (4.18) \\
&\leq T_B + 1 + \sum_{t\geq T_B+1} \exp\left(\frac{-c_{\min}^2 t}{2}\right) \qquad (4.19) \\
&\leq T_B + 1 + \frac{2}{c_{\min}^2} \exp\left(\frac{c_{\min}^2}{2} - \frac{c_{\min}^2 (T_B+1)}{2}\right) \qquad (4.20) \\
&\leq T_B + 1 + \frac{2}{c_{\min}^2} \exp(-c_{\min} B),
\end{aligned}$$

where (4.18) makes use of Fact 2, (4.19) is obtained because $2(B - tc_{\min})^2 \geq c_{\min}^2 t^2/2$ for $t \geq 2B/c_{\min}$ and we get (4.20) since $1/(1 - e^{-c_{\min}^2/2}) \leq 2e^{c_{\min}^2/2}/c_{\min}^2$.

Fact 2 (Hoeffding, 1963; Flajolet and Jaillet, 2015)**.** *Let $X_1, \ldots, X_t$ be the random variables with common support $[0,1]$ and such that there exists $a \in \mathbb{R}$ with $\forall u \in [t]$, $\mathbb{E}[X_u | X_1, \ldots, X_{u-1}] \geq a$. Let $\overline{x}_t \triangleq \frac{1}{t}(X_1 + \cdots + X_t)$, then*

$$\forall \varepsilon \geq 0 \ , \ \mathbb{P}[\overline{x}_t - a \leq -\varepsilon] \leq e^{-2\varepsilon^2 t}.$$

The above proved upper bound on the expected horizon is sufficient for our purpose, as we are interested in the asymptotic rate of regret when B is large. We can nevertheless note that in the regime where $c_{\min}$ is small, our gap dependent bound has only a logarithmic dependence in $c_{\min}^{-1}$. □

4.5.2 Proof of Theorem 20

Proof of Theorem 20. Let $0 < \varepsilon < 1/4$. We consider a DAG composed of two disjoint paths (Figure 4.1), both with $n/2$ nodes. We denote the two paths by S and S'. We deterministically set all the costs to 1, $w_i = 0$ for $i \notin \left\{s_{\frac{n}{2}}, s'_{\frac{n}{2}}\right\}$. All this information is given to the agent. Notice that this does not make the problem harder.

Now consider two distributions D_1 and D_2 defined by

$$D_1: \quad w_{s_{\frac{n}{2}}} \triangleq \frac{1}{2} + \varepsilon, \quad w_{s'_{\frac{n}{2}}} \triangleq \frac{1}{2} - \varepsilon \quad \text{and} \quad D_2: \quad w_{s_{\frac{n}{2}}} \triangleq \frac{1}{2} - \varepsilon, \quad w_{s'_{\frac{n}{2}}} \triangleq \frac{1}{2} + \varepsilon.$$

Notice that an optimal online policy does not modify its behavior during a round t, since after having seen $W_{i,t} = 1$, continuing searching would only give information about cost distribution which is known by the problem definition, and no additional

information about the rewards. Therefore, there is *an optimal* online policy that selects some search S_t and perform A_t over round t. Observe that $S^* = SS'$ for D_1 and $S^* = S'S$ for D_2. We have $J^* = \frac{3}{4}n - \varepsilon n \geq \frac{1}{2}n$ for both D_1 and D_2.

We now show that we can restrict ourselves to policies that take searches in $\{SS', S'S\}$.

- First, an optimal online policy does not select a search that would not include at least one of the leaves $\left\{s_{\frac{n}{2}}, s'_{\frac{n}{2}}\right\}$ for a round. Therefore, it has a full information on W. Indeed, a search of support included in $\left\{s_{\frac{n}{2}}, s'_{\frac{n}{2}}\right\}^c$ is noninformative and does not bring any reward while having a cost.

- Second, for a policy π that does not select a search in $\{SS', S'S\}$ for some round t, we construct π' that acts like π except for this round t where it selects SS' if π would see the leaf $s_{\frac{n}{2}}$ first, and $S'S$ otherwise, i.e., if π would first see the leaf $s_{\frac{n}{2}}$. Now compare both policies on the same realization of $\mathbf{W}_1, \mathbf{W}_2, \ldots$. We claim that the global reward of π' is never smaller than that of π. By symmetry, assume that π sees $s_{\frac{n}{2}}$ first within round t and thus π' selects SS'.

  - If $W_{s_{\frac{n}{2}},t} = 1$ or $\left(W_{s'_{\frac{n}{2}},t} = 1 \text{ and } \pi \text{ visits } s'_{\frac{n}{2}} \text{ within round } t\right)$, both policies obtain the same reward of 1 within round t, but π' pays less than π.
  - If $W_{s'_{\frac{n}{2}},t} = 1$ and π does not visit $s'_{\frac{n}{2}}$ within round t, π gains 0 and pays at least $n/2$, whereas π' gains 1 and pays n within round t. Thus, the budget of π compared to π' is augmented by at most $n/2$, with which it can increase its reward by at most 1.

  The overall reward of π' remains higher than that of π for both cases.

A direct consequence of the restriction to $\{SS', S'S\}$ is that $c_{\min} = n/2$, giving the upper bound in Theorem 20 by invoking the result of Theorem 19.

Now for a policy π using searches from $\{SS', S'S\}$, we have

$$
\begin{aligned}
F_B(\pi) &= \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S_t} \mathbb{I}\{B_t \geq 0,\ W_{i,t} = 1\}\right] \\
&\leq \sum_{t=1}^{\infty} \mathbb{E}\left[\sum_{i \in S_t} \mathbb{I}\{B_{t-1} \geq 0,\ W_{i,t} = 1\}\right] \\
&= \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \in \mathcal{S}^*\} \mathbf{e}_{S_t}^\intercal \mathbf{w}^*\right] + \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\} \mathbf{e}_{S_t}^\intercal \mathbf{w}^*\right] \\
&= \frac{1}{J^*} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \in \mathcal{S}^*\} \left(d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{e}_{S_t}^\intercal \mathbf{w}^*) \mathbf{e}_{S_t}^\intercal \mathbf{c}^*\right)\right] \\
&\quad + \frac{1}{J^*} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\} \left(d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{e}_{S_t}^\intercal \mathbf{w}^*) \mathbf{e}_{S_t}^\intercal \mathbf{c}^*\right)\right] \\
&\quad - \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\} \Delta(S_t)\right] \\
&= \frac{1}{J^*} \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0\} \left(d(S_t; \mathbf{w}^*, \mathbf{c}^*) + (1 - \mathbf{e}_{S_t}^\intercal \mathbf{w}^*) \mathbf{e}_{S_t}^\intercal \mathbf{c}^*\right)\right] \\
&\quad - \sum_{t \geq 1} \mathbb{E}\left[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\} \Delta(S_t)\right]
\end{aligned}
$$

$$\leq \frac{B+n}{J^*} - \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\}\Delta(S_t)].$$

As a result we get

$$\begin{aligned} B_B(\pi) = F_B^* - F_B(\pi) &\geq \frac{B-n}{J^*} - \frac{B+n}{J^*} + \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\}\Delta(S_t)] \\ &= -\frac{2n}{J^*} + \sum_{t\geq 1} \mathbb{E}[\mathbb{I}\{B_{t-1} \geq 0,\ S_t \notin \mathcal{S}^*\}\Delta(S_t)]\cdot \end{aligned}$$

Since we restrict π to take a search in $\{SS', S'S\}$, we have a single gap (the same for D_1 and D_2):

$$\Delta = \frac{\frac{n}{2}\left(\frac{1}{2}-\varepsilon\right) + n\left(\frac{1}{2}+\varepsilon\right)}{\frac{n}{2}\left(\frac{1}{2}+\varepsilon\right) + n\left(\frac{1}{2}-\varepsilon\right)} - 1 = \frac{1.5+\varepsilon}{1.5-\varepsilon} - 1 = \frac{2\varepsilon}{1.5-\varepsilon} \geq \frac{4\varepsilon}{3}\cdot \tag{4.21}$$

Furthermore we can bound the number of rounds from below by B/n. To proceed we use high-probability Pinsker inequality Tsybakov, 2009, Lemma 2.6.

Fact 3 (high-probability Pinsker inequality)**.** *Let P and Q be probability measures on the same measurable space, and let $\mathfrak{A}$ be an event. Then*

$$P(\mathfrak{A}) + Q(\neg\mathfrak{A}) \geq \frac{1}{2}\exp(-\mathrm{KL}(P\|Q)),$$

where KL is the Kullback-Leibler divergence.

We let $R_{1,B}(\pi)$ be the regret of π for distribution D_1 and similarly, $R_{2,B}(\pi)$ for D_2. If $\mathbb{P}_1$ and $\mathbb{P}_2$ denote the probability when random variable are samples from D_1 and D_2 respectively, we have

$$\begin{aligned} &\max\{R_{1,B}(\pi), R_{2,B}(\pi)\} \\ &\geq \frac{R_{1,B}(\pi) + R_{2,B}(\pi)}{2} \qquad (4.22) \\ &\geq -\frac{2n}{J^*} + \frac{\Delta}{2}\sum_{t=1}^{B/n}(\mathbb{P}_1[B_{t-1}\geq 0, S_t = S'S] + \mathbb{P}_2[B_{t-1}\geq 0, S_t = SS']) \\ &\geq -\frac{2n}{J^*} + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp\left(-\mathrm{KL}\left(D_1^{\otimes t}\|D_2^{\otimes t}\right)\right) \qquad (4.23) \\ &= -\frac{2n}{J^*} + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp(-t\mathrm{KL}(D_1\|D_2)), \end{aligned}$$

where (4.23) is due to Fact 3 and (4.21). Then,

$$\begin{aligned} \mathrm{KL}(D_1\|D_2) &= \left(\frac{1}{2}+\varepsilon\right)\log\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}\right) + \left(\frac{1}{2}-\varepsilon\right)\log\left(\frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}+\varepsilon}\right) \\ &\leq 2\varepsilon\left(\frac{\frac{1}{2}+\varepsilon}{\frac{1}{2}-\varepsilon}\right) - 2\varepsilon\left(\frac{\frac{1}{2}-\varepsilon}{\frac{1}{2}+\varepsilon}\right) = \frac{4\varepsilon^2}{\frac{1}{4}-\varepsilon^2} \leq \frac{64}{3}\varepsilon^2 \qquad \log(x)\leq x-1. \end{aligned}$$

Thus, with $J^* \geq \frac{n}{2}$, we have

$$\begin{aligned}\max\{R_{1,B}(\pi), R_{2,B}(\pi)\} &\geq -4 + \frac{\varepsilon}{3}\sum_{t=1}^{B/n}\exp\Big(-\frac{64}{3}t\varepsilon^2\Big) \geq -4 + \frac{\varepsilon\Big(1-\exp\Big(-\frac{64}{3}\frac{B\varepsilon^2}{n}\Big)\Big)}{3\Big(\exp\Big(\frac{64}{3}\varepsilon^2\Big)-1\Big)} \\ &\geq -4 + \frac{1-\exp\Big(-\frac{64}{3}\frac{B\varepsilon^2}{n}\Big)}{64\varepsilon} \\ &\geq -4 + \min\Big\{\frac{1}{128\varepsilon}, \frac{\varepsilon B}{6n}\Big\}.\end{aligned}$$

Taking $\varepsilon = \sqrt{(6n)/(128B)}$, the lower bound becomes

$$\max\{R_{1,B}(\pi), R_{2,B}(\pi)\} \geq -4 + \sqrt{\frac{B}{768n}} \geq -4 + \frac{1}{28}\sqrt{\frac{B}{n}}.$$

□

Chapter 5

The Structure of Uncertainty

This chapter aims to introduce some tools necessary to the construction of more sophisticated confidence regions than those we saw in the previous chapter. We will then see that although the use of these confidence regions can lead to improved regret bounds, they also can lead to efficiency issue in the algorithms, and that in some cases, more complex optimization methods have to be implemented. The beginning of the chapter is mainly based on known structured bandit results. More precisely, we will rely on Degenne and Perchet (2016b) and on Magureanu, Combes, and Proutiere (2014). The end of the chapter is based on our article Perrault, Perchet, and Valko (2019a).

About confidence regions Here we give a more precise idea of the type of confidence region we are interested in. In the previous chapter, we could see that the optimistic estimates of the means were constructed from either the Hoeffding inequality or the Bernstein's one (in the end, we also mentioned the use of kl-based confidence intervals). In other words, if we look at the confidence regions from the vectorial point of view, they all have the shape of a hypercube (i.e., a Cartesian product of confidence intervals). Note that these confidence regions are all from Chapter 2, i.e., they are all already used for the standard MAB setting. Here we want to investigate other types of confidence regions, which will be based more on the ℓ_2-norm (ellipsoid) rather than the ℓ_∞-norm (hypercube). Indeed, we saw in the previous chapter that a region based on the ℓ_∞-norm induces the use of bonuses of type ℓ_1. We'll see that, naturally, the use of regions based on the ℓ_2-norm will induce the use of ℓ_2-type bonuses. This is interesting because, as we saw in Chapter 3, there is a gain in using the ℓ_2-bonuses compared to the ℓ_1-bonuses. For example, in the case where $A_t = S_t$ is of maximum cardinality m, $b_i(S_t) = 1$ and $\alpha_i = 1/2$ (note that these settings are the most standard to consider), the use of an ℓ_2-bonus improves the regret by a factor $m \log^{-2}(m)$ (more precisely, the linear dependence of the regret in the batch size m can be replaced by a polylogarithmic dependence $\log^2(m)$). In the following, we will see two methods for building ellipsoidal confidence regions. Of course, these constructions (and the resulting gain in the regret bound) are not free. Specifically, the ℓ_2-regions are generally based on an assumption about the outcomes $\mathbf{X}$. As a reminder, the assumptions for the ℓ_∞-regions only concern the marginals $\mathbb{P}_{X_i}$ (like the κ^2-sub-Gaussianity, the boundedness, ...), and nothing is said about the type of dependence that exists between these marginals. So, by providing additional information about these dependencies, we get more *structure* about the uncertainty associated with the outcomes $\mathbf{X}$. This structure of uncertainty can be exploited in order to improve the performance of the policies. The most natural way to proceed, which we will follow here for sub-Gaussianity, is to extend the structures used for the marginals to the multivariate case.

For the two methods we look at, we make the following assumption on the vector of outcomes $\mathbf{X}$.

Assumption 1 (Subgaussian outcomes)**.** *There is a matrix* $\mathbf{C} \succeq 0$ *and a* known *matrix* $\mathbf{\Gamma} \succeq_+ \mathbf{C}$ *with* $\Gamma_{ij} \in \mathbb{R}_+$ *for all* $i, j \in [n]$*, such that*

$$\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal \mathbf{C} \boldsymbol{\lambda}}.$$

The motivation for sub-Gaussian outcomes is the following: 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 $\mathcal{K}$.

Here, we focus on the linear reward case, for sake of simplicity, with an optimal action denoted A^* (so that $\Delta_t = \boldsymbol{\mu}^{*\intercal}(\mathbf{e}_{A^*} - \mathbf{e}_{A_t})$). This construction will inspire us to generalize it to a more complex reward function. A common statistic computed from the feedback received up to the beginning of round t is the *empirical average* of each arm $i \in [n]$, defined as

$$\overline{\mu}_{i,0} = 0, \quad \forall t \geq 2,\ \overline{\mu}_{i,t-1} = \frac{\sum_{t' \in [t-1]} \mathbb{I}\{i \in A_{t'}\} X_{i,t'}}{N_{i,t-1}}$$

where $\forall t \geq 1,\ N_{i,t-1} \triangleq \sum_{t' \in [t-1]} \mathbb{I}\{i \in A_{t'}\}$.

5.1 Laplace's method

We review here the most used method for ellipsoidal confidence region for $\boldsymbol{\mu}^*$: the Laplace's method. A very detailed explanation of Laplace's method is given in Lattimore and Szepesvári (2019) and in Abbasi-Yadkori, Pál, and Szepesvári (2011). Here, we are going to focus on its use in our context of semi-bandit feedback. The difficulty of designing confidence sets for $\boldsymbol{\mu}^*$ stems from the fact that the actions are correlated to the outcomes (because the agent is adaptive).

Since we are considering a linear reward function, we want to compare $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}_{t-1}$ in some fixed direction $\mathbf{e}_A$, where $A \in \mathcal{A}$. Indeed, we can first obtain the following proposition in the same way as Theorem 9.

Proposition 14. *The regret under the event* $(\overline{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*)^\intercal \mathbf{e}_{A_t} > \Delta_t/2$ *is bounded by* $\Delta_{\max} 8nm^2 \max_i C_{ii} / \Delta_{\min}^2$.

Proof. Using Assumption 1,

$$\begin{aligned}
&\sum_{t=1}^{T} \Delta_t \mathbb{P}\left[(\overline{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*)^\intercal \mathbf{e}_{A_t} > \Delta_t/2\right] \\
&\leq \sum_{t=1}^{T} \Delta_t \sum_{i \in A_t} \mathbb{P}\left[\overline{\mu}_{i,t-1} - \mu_i^* > \Delta_{\min}/(2m)\right] \\
&\leq \sum_{t=1}^{T} \Delta_{\max} \sum_{i \in [n]} \mathbb{P}\left[i \in A_t,\ \overline{\mu}_{i,t-1} - \mu_i^* > \Delta_{\min}/(2m)\right] \\
&\leq \Delta_{\max} \sum_{i \in [n]} \sum_{t \in [T]} e^{-\Delta_{\min}^2 t/(8m^2 C_{ii})} \leq \frac{\Delta_{\max} 8nm^2 \max_i C_{ii}}{\Delta_{\min}^2}.
\end{aligned}$$

□

We can thus safely assume that $(\overline{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*)^\mathsf{T}\mathbf{e}_{A_t} \leq \Delta_t/2$. Then, bounding $\mathbf{e}_{A^*}^\mathsf{T}(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1})$ by some $\text{bonus}_t(A^*)$ with high probability, and using a policy that chooses A_t maximizing $A \mapsto \overline{\boldsymbol{\mu}}_{t-1}^\mathsf{T}\mathbf{e}_A + \text{bonus}_t(A)$, we have

$$\begin{aligned}\Delta_t &= \boldsymbol{\mu}^{*\mathsf{T}}(\mathbf{e}_{A^*} - \mathbf{e}_{A_t}) \\ &\leq \overline{\boldsymbol{\mu}}_{t-1}^\mathsf{T}\mathbf{e}_{A^*} + \text{bonus}_t(A^*) - \boldsymbol{\mu}^{*\mathsf{T}}\mathbf{e}_{A_t} \\ &\leq \overline{\boldsymbol{\mu}}_{t-1}^\mathsf{T}\mathbf{e}_{A_t} + \text{bonus}_t(A_t) - \boldsymbol{\mu}^{*\mathsf{T}}\mathbf{e}_{A_t} \\ &\leq \text{bonus}_t(A_t) + \Delta_t/2,\end{aligned}$$

This provides the bound $\Delta_t \leq 2\text{bonus}_t(A_t)$, which is what we want to apply theorems from the end of Chapter 3. The bound on $\mathbf{e}_{A^*}^\mathsf{T}(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1})$ is obtained with the following Theorem 21.

Theorem 21 (Laplace's method)**.** *For any $A \in \mathcal{A}$, with probability at least $1 - \frac{1}{t\log^2(t)} \wedge 1$, we have that $\mathbf{e}_A^\mathsf{T}(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1})$ is upper bounded by*

$$\sqrt{2\delta(t)}\left\|\mathbf{e}_A^\mathsf{T}\text{diag}\left(\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}}\right)^{-1}\right\|_{\text{diag}\left(\eta\Gamma_{ii}\sum_{t'=1}^{t-1}\mathbb{I}\{i\in A_{t'}\}\right)_i + \sum_{t'=1}^{t-1}\boldsymbol{\Gamma}\odot\left(\mathbf{e}_{A_{t'}}\mathbf{e}_{A_{t'}}^\mathsf{T}\right)},$$

for some real parameter $\eta > 0$, and where $\delta(t) = \log(t) + (2+m)\log\log(t) + m\log(1+e/\eta)/2$.

Proof. We can write the following using the Cauchy-Schwarz inequality.

$$\begin{aligned}\mathbf{e}_A^\mathsf{T}(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1}) &= \mathbf{e}_A^\mathsf{T}\text{diag}\left(\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\right)^{-1}\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^* - \mathbf{X}_{t'}) \\ &\leq \left\|\mathbf{e}_A^\mathsf{T}\text{diag}\left(\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\right)^{-1}\right\|_{\mathbf{C}_{t-1}}\left\|\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^* - \mathbf{X}_{t'})\right\|_{\mathbf{C}_{t-1}^{-1}},\end{aligned}$$

where the matrix $\mathbf{C}_{t-1}$ is $D + \sum_{t'=1}^{t-1}\mathbf{C}\odot\left(\mathbf{e}_{A_{t'}\cap A}\mathbf{e}_{A_{t'}\cap A}^\mathsf{T}\right)$, with D being a positive definite diagonal matrix to be specified latter. Our goal is to bound the right hand term with high probability. This is an interesting quantity for the following reason:

$$\frac{1}{2}\left\|\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^* - \mathbf{X}_{t'})\right\|_{\mathbf{C}_{t-1}^{-1}}^2 = \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} -\frac{1}{2}\boldsymbol{\lambda}^\mathsf{T}\mathbf{C}_{t-1}\boldsymbol{\lambda} + \boldsymbol{\lambda}^\mathsf{T}\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^* - \mathbf{X}_{t'}),$$

and composing by the exponential, we begin to see a form similar to the sub-Gaussianity assumption mentioned above. More precisely, for a fixed $\boldsymbol{\lambda}$, the expectation of the exponential of the RHS is lower than 1. However, we see here that there is a maximum over $\boldsymbol{\lambda}$. The Laplace's method (also called method of mixtures) is precisely a way to get rid of this maximum. Intuitively, it provides an approximation of the maximum by integrating the exponential against a multivariate normal centered at $\mathbf{C}_{t-1}^{-1}\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^* - \mathbf{X}_{t'})$, where the maximum is attained. The integrals over $\boldsymbol{\lambda}$ and the expectation can then be swapped by Fubini's theorem, to get an approximation of the expectation of the maximum using an integral of the expectations: Let f be the density of a multivariate normal random variable independent of all other variables with mean 0 and covariance D^{-1}, we have

$$\sqrt{\frac{\det(D)}{\det(\mathbf{C}_{t-1})}}\exp\left(\frac{1}{2}\left\|\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^*-\mathbf{X}_{t'})\right\|^2_{\mathbf{C}_{t-1}^{-1}}\right)$$
$$=\int_{\boldsymbol{\lambda}\in\mathbb{R}^n}\prod_{t'=1}^{t-1}\exp\Big(-\frac{1}{2}\boldsymbol{\lambda}^\intercal\mathbf{C}\odot\Big(\mathbf{e}_{A_{t'}\cap A}\mathbf{e}_{A_{t'}\cap A}^\intercal\Big)\boldsymbol{\lambda}+\boldsymbol{\lambda}^\intercal\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^*-\mathbf{X}_{t'})\Big)f(\boldsymbol{\lambda})\mathrm{d}\boldsymbol{\lambda}.$$

From Assumption 1, we see that the product inside the integral is a supermartingale (with respect to $\mathcal{F}_{t-1}$). Thus, taking the expectation, we have

$$\mathbb{E}\left[\sqrt{\frac{\det(D)}{\det(\mathbf{C}_{t-1})}}\exp\left(\frac{1}{2}\left\|\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^*-\mathbf{X}_{t'})\right\|^2_{\mathbf{C}_{t-1}^{-1}}\right)\right]\leq 1.$$

To deal with $\sqrt{\frac{\det(D)}{\det(\mathbf{C}_{t-1})}}$, we use a peeling argument: we consider the event under which for all $i\in A$, $e^{a_i}\leq N_{i,t-1}\leq e^{a_i+1}$. The number of $\mathbf{a}$ to consider to cover all the possibilities is $\log(t)^m$. With the choice $D_i=\mathbb{I}\{i\in A\}\eta e^{a_i}C_{ii}+\mathbb{I}\{i\notin A\}$, we get the lower bound

$$\sqrt{\frac{\det(D)}{\det(\mathbf{C}_{t-1})}}\geq(1+e/\eta)^m.$$

In summary, we have using Markov's inequality that

$$\mathbb{P}\left[\left\|\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\odot(\boldsymbol{\mu}^*-\mathbf{X}_{t'})\right\|_{\mathbf{C}_{t-1}^{-1}}\geq\delta(t)\right]\leq\frac{1}{t\log^2(t)}\wedge 1,$$

with $\delta(t)=\log(t)+(2+m)\log\log(t)+m\log(1+e/\eta)/2$. The proof is concluded using the assumption that $\mathbf{C}\succeq_+\boldsymbol{\Gamma}$. □

The Laplace's method was originally used for linear bandits (Abbasi-Yadkori, Pál, and Szepesvári, 2011). The above proof in taken from Degenne and Perchet (2016b), who adapted this method to the semi-bandit feedback setting. Bounds like those given in Theorem 21 are called self-normalized bounds (Peña, Lai, and Shao, 2008), because the control is with respect to the norm associated with the inverse of a regularized design matrix. We notice that this matrix is non-diagonal if $\boldsymbol{\Gamma}$ is, and that we do not have in our arsenal a theorem dealing with bonus of this form. To the best of our knowledge, there is no such work that treats directly this kind of bonus for semi-bandit feedback, which represents an interesting open question.

A first way to proceed is to further bound this bonus by an ℓ_2-type one, i.e., corresponding to a diagonal matrix (the associated confidence region is then ellipsoidal, but is aligned with respect to the axes). More precisely, if we define the non-diagonal matrix $\mathbf{M}'=\sum_{t'=1}^{t-1}\boldsymbol{\Gamma}\odot\left(\mathbf{e}_{A_{t'}}\mathbf{e}_{A_{t'}}^\intercal\right)$, we can rely on the inequality

$$\left\|\mathbf{e}_A^\intercal\mathrm{diag}\left(\sum_{t'=1}^{t-1}\mathbf{e}_{A_{t'}\cap A}\right)^{-1}\right\|^2_{\mathbf{M}'}=\sum_{i,j\in A}\frac{N_{ij,t-1}\Gamma_{ij}}{N_{i,t-1}N_{j,t-1}}$$
$$\leq\sum_{i\in A}\frac{1}{N_{i,t-1}}\sum_{j\in A}\Gamma_{ij}\leq\sum_{i\in A}\frac{1}{N_{i,t-1}}\max_{A'\in\mathcal{A},\,i\in A'}\sum_{j\in A'}\Gamma_{ij},$$

where $N_{ij,t-1} = \sum_{t'=1}^{t-1} \mathbb{I}\{i,j \in A_{t'}\}$.

A second method, which has been used by Degenne and Perchet (2016b), is to decompose the bonus as

$$\left\| \mathbf{e}_A^\intercal \mathrm{diag}\left(\sum_{t'=1}^{t-1} \mathbf{e}_{A_{t'} \cap A} \right)^{-1} \right\|_{\mathbf{M}'}^2 \leq (1-\gamma) \sum_{i \in A} \frac{\Gamma_{ii}}{N_{i,t-1}} + \gamma \left(\sum_{i \in A} \sqrt{\frac{\Gamma_{ii}}{N_{i,t-1}}} \right)^2, \tag{5.1}$$

where $\gamma \triangleq \max_{(i,j) \in A^2, i \neq j} \Gamma_{ij} / \sqrt{\Gamma_{ii}\Gamma_{jj}}$. We thus get this way a decomposed bonus (with an ℓ_1 term and an ℓ_2 one).

We want to insist on the fact that the original bonus can be used in the algorithm, and that these two processes for changing the bonus are only used in the analysis.

5.2 Covering argument

An alternative to the method previously proposed is to use a *covering argument*[1]. We saw that the main issue is that the max and the expectation couldn't be swapped. Put it another way, the issue is that $\boldsymbol{\lambda}$ is random when we want to use the sub-Gaussianity inequality. We have a clear expression for $\boldsymbol{\lambda}$, and the randomness comes mainly from the empirical mean (also from counters, but these are treated separately using a peeling). A covering argument is another approach that removes the randomness in $\boldsymbol{\lambda}$ by replacing, in its definition, the empirical mean by a deterministic quantity that is provably close to it. A positive point of the covering method is that it can be easily extended to a more general setting than the sub-Gaussian one, as we will see in Chapter 7. A negative point is that it is tricky to use when the matrix $\boldsymbol{\Gamma}$ is not diagonal. Thus, in this section, we take $\boldsymbol{\Gamma}$ diagonal in Assumption 1 (notice, this happens for instance when outcomes are mutually independent and component-wise sub-Gaussian).

The following Theorem 22 is adapted from Magureanu, Combes, and Proutiere (2014). In particular, under the complementary event considered in the theorem, we can use the Cauchy-Schwarz inequality to get a high probability bound on

$$\mathbf{e}_A^\intercal(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1}) \leq \mathbf{e}_A^\intercal(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1}) \vee 0$$

that has the form of an ℓ_2-bonus.

Theorem 22 (Covering argument)**.** *In the case* $\boldsymbol{\Gamma}$ *is diagonal, we have*

$$\mathbb{P}\left[\sum_{i \in A} \mathbb{I}\left\{ \mu_i^* \geq \overline{\mu}_{i,t-1} \right\} N_{i,t-1} \frac{\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)^2}{2\Gamma_{ii}} \geq \delta \right] \leq e^{m+1} \left(\frac{(\delta-1)\log(t)}{m} \right)^m e^{-\delta},$$

for some $\delta \geq m+1$.

Proof. We fix some $\delta \geq m+1$, and define the following events:

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

[1] Although the method presented is quite different from the classical "covering argument" used in linear bandits, we will still call this method so, in order to distinguish it from the Laplace method.

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

Since each number of pulls $N_{i,t-1}$ for $i \in A$ is bounded by t, the number of possible $\mathbf{d} \in (\mathbb{N}^*)^A$ such that $\mathbb{P}[\mathfrak{B}_{\mathbf{d},t}] > 0$ is bounded by $(\log(t)/\log(\delta/(\delta-1)))^m$. Thanks to the following Lemma 6, and an union bound on those possible $\mathbf{d} \in (\mathbb{N}^*)^A$, we get

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

Lemma 6. *Let* $\mathbf{d} \in (\mathbb{N}^*)^A$. *Then,* $\mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{\mathbf{d},t}] \leq \left(\frac{(\delta-1)e}{m}\right)^m e^{1-\delta}$.

Proof. The idea is to get rid of randomness by replacing the empirical mean $\overline{\mu}_{i,t-1}$ by some non-random value x_i. Let $\boldsymbol{\zeta} \in \mathbb{R}_+^A$. For $i \in A$, we define $x_i(N) = \mu_i^* - \sqrt{\frac{2\zeta_i \Gamma_{ii}}{N}}$. Under the events $\mathfrak{B}_{\mathbf{d},t}$ and

$$\mathfrak{A}'_t \triangleq \bigcap_{i \in A} \left\{ N_{i,t-1} \frac{\left(0 \vee \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)^2}{2\Gamma_{ii}} > \zeta_i \right\},$$

we have

$$\bigcap_{i \in A} \left\{ \overline{\mu}_{i,t-1} \leq x_i \left(\left(\frac{\delta}{\delta-1}\right)^{d_i} \right) \right\}. \tag{5.2}$$

With $\varepsilon_i \triangleq \mu_i^* - x_i\left(\left(\frac{\delta}{\delta-1}\right)^{d_i}\right)$ and $\lambda_i \triangleq \frac{\varepsilon_i}{\Gamma_{ii}}$, $i \in A$, this further implies:

$$\begin{aligned}
\frac{\delta-1}{\delta} \sum_{i \in A} \zeta_i &= \sum_{i \in A} \left(\frac{\delta}{\delta-1}\right)^{d_i-1} \frac{\varepsilon_i^2}{2\Gamma_{ii}} \\
&\leq \sum_{i \in A} N_{i,t-1} \frac{\varepsilon_i^2}{2\Gamma_{ii}} && \mathfrak{B}_{\mathbf{d},t} \\
&= \sum_{i \in A} N_{i,t-1} \lambda_i \varepsilon_i - \sum_{i \in A} N_{i,t-1} \Gamma_{ii} \lambda_i^2/2 \\
&\leq \sum_{i \in A} N_{i,t-1} \lambda_i \left(\mu_i^* - \overline{\mu}_{i,t-1}\right) - \sum_{i \in A} N_{i,t-1} \Gamma_{ii} \lambda_i^2/2 && \text{using (5.2).}
\end{aligned}$$

This last quantity can be rewritten as

$$\sum_{t' \in [t-1]} \left(\left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal (\boldsymbol{\mu}^* - \mathbf{X}_{t'}) - \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal \boldsymbol{\Gamma} \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)/2 \right).$$

Since $\boldsymbol{\lambda} \geq 0$, we have using Assumption 1

$$\left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal \mathbf{C} \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right) \leq \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal \boldsymbol{\Gamma} \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right),$$

Thus, still using Assumption 1, we have

$$\frac{\delta-1}{\delta} \sum_{i \in A} \zeta_i \leq \sum_{t' \in [t-1]} \left(\left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal (\boldsymbol{\mu}^* - \mathbf{X}_{t'}) - \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)^\intercal \mathbf{C} \left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A}\right)/2 \right)$$

$$\leq \sum_{t' \in [t-1]} \left(\left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A} \right)^\mathsf{T} (\boldsymbol{\mu}^* - \mathbf{X}_{t'}) - \log \mathbb{E} \left[e^{\left(\boldsymbol{\lambda} \odot \mathbf{e}_{A_{t'} \cap A} \right)^\mathsf{T} (\mathbf{X} - \boldsymbol{\mu}^*)} \right] \right).$$

The proof now follows the usual path, and we get using Markov's inequality that

$$\mathbb{P} \left[\bigcap_{i \in A} \left\{ \mathbb{I}\{\mathfrak{B}_{\mathbf{d},t}\} N_{i,t-1} \frac{\left(0 \vee \left(\mu_i^* - \overline{\mu}_{i,t-1} \right) \right)^2}{2\Gamma_{ii}} > \zeta_i \right\} \right] \leq e^{-\sum_{i \in A} \zeta_i \frac{\delta - 1}{\delta}}.$$

By Lemma 7 (Lemma 8 of Magureanu, Combes, and Proutiere (2014)), since $\delta \geq m + 1$, we have

$$\mathbb{P}[\mathfrak{B}_{\mathbf{d},t} \cap \mathfrak{A}_t] \leq \left(\frac{(\delta - 1)e}{m} \right)^m e^{1-\delta}.$$

□

Lemma 7. *For $K \geq 2$, $a > 0$, $\mathbf{Z} \in \mathbb{R}^K$ a random vector and $\boldsymbol{\zeta} \in \mathbb{R}_+^K$ a vector, if $\mathbb{P}[\mathbf{Z} \geq \boldsymbol{\zeta}] \leq e^{-a \sum_i \zeta_i}$, then for $\delta \geq K/a$,*

$$\mathbb{P} \left[\sum_i Z_i \geq \delta \right] \leq \left(\frac{a\delta e}{K} \right)^K e^{-a\delta}.$$

Proof. We have with $\lambda = a - K/\delta \geq 0$

$$\begin{aligned}
\mathbb{P} \left[\sum_i Z_i \geq \delta \right] &\leq \mathbb{E} \left[\prod_i e^{\lambda Z_i} \right] e^{-\lambda\delta} \\
&= \int_{\mathbb{R}_+^K} \mathbb{P} \left[e^{\lambda \mathbf{Z}} > \mathbf{x} \right] \mathrm{d}\mathbf{x} e^{-\lambda\delta} \\
&= \int_{\mathbb{R}_+^K} \mathbb{P}[\mathbf{Z} > \log(\mathbf{x})/\lambda] \mathrm{d}\mathbf{x} e^{-\lambda\delta} \\
&\leq \left(1 + \int_1^\infty e^{-a \log(x)/\lambda} \mathrm{d}x \right)^K e^{-\lambda\delta} \\
&= \left(\frac{a}{a - \lambda} \right)^K e^{-\lambda\delta} = \left(\frac{a\delta e}{K} \right)^K e^{-a\delta}.
\end{aligned}$$

□

□

5.3 Efficiency of algorithms

In the two previous sections, we have seen how concentration inequalities could be derived to build ℓ_2-bonuses. Then, standard optimistic policies (e.g., ESCB in Combes et al. (2015) and OLS-UCB in Degenne and Perchet (2016b)) chose action A_t that solves the following combinatorial optimization problem at each round t (in the case of a linear reward function):

$$\max_{A \in \mathcal{A}} \mathbf{e}_A^\mathsf{T} \overline{\boldsymbol{\mu}}_{t-1} + \mathrm{bonus}_t(A), \tag{5.3}$$

where $\mathrm{bonus}_t(A)$ is the ℓ_2-bonus, i.e., one of the high probability upper bounds on $\mathbf{e}_A^\mathsf{T}(\boldsymbol{\mu}^* - \overline{\boldsymbol{\mu}}_{t-1})$ we considered in the previous sections.

Class of possible outcome distributions	*Gap-dependent lower bound*	*Gap-dependent upper bound*
$(i)+(iii)$ $\Rightarrow(i)+(v)$ $\Rightarrow(iv)$	$\Omega\left(\frac{n\log T}{\Delta}\right)$	ℓ_2-bonus: $\mathcal{O}\left(\frac{n\log^2(m)\log T}{\Delta}\right)$
$(ii)+(iii)$ $\Rightarrow(ii)+(v)$	$\Omega\left(\frac{nm\log T}{\Delta}\right)$	ℓ_1-bonus: $\mathcal{O}\left(\frac{nm\log T}{\Delta}\right)$

TABLE 5.1: Gap-dependent lower bounds proved on different classes of possible distributions for $\mathbf{X}$.

In what follows, we give a more precise overview of the improvement induced by the use of ℓ_2-bonuses compared to ℓ_1-bonuses, stressing that this improvement is unavoidable since the two approaches are matching two different lower bounds. Next, we identify the inefficiency of the ℓ_2-bonus policies by noting that for many situations that are considered in CMAB, the above combinatorial optimization problem (5.3) is not tractable. Finally, we propose an efficient implementation in the case $\mathcal{A}$ has a *matroid* structure.

5.3.1 Lower bounds

In an instance of a CMAB problem, we can consider different properties satisfied by the outcomes $\mathbf{X}$. For example, some properties that are commonly assumed are:

(i) $X_1,\dots,X_n\in\mathbb{R}$ are *mutually independent*,

(ii) $X_1,\dots,X_n\in\mathbb{R}$ are *arbitrary correlated*,

(iii) $\mathbf{X}\in[-1,1]^n$,

(iv) $\mathbf{X}\in\mathbb{R}^n$ is *multivariate sub-Gaussian*,
i.e., $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right]\leq e^{\|\boldsymbol{\lambda}\|_2^2/2},\ \forall\boldsymbol{\lambda}\in\mathbb{R}^n$,

(v) $\mathbf{X}\in\mathbb{R}^n$ is *component-wise sub-Gaussian*,
i.e., $\mathbb{E}\left[e^{\lambda_i\left(X_i-\mu_i^*\right)}\right]\leq e^{\lambda_i^2/2},\ \forall i\in[n],\ \forall\boldsymbol{\lambda}\in\mathbb{R}^n$.

Combining some of the above properties, we consider different classes of possible distributions for $\mathbf{X}$. In the following, we consider five examples of such combination (classifying them by implication):

- $(i)+(iii)\Rightarrow(i)+(v)\Rightarrow(iv)$,
- $(ii)+(iii)\Rightarrow(ii)+(v)$.

Notice, the above classes are special cases of Assumption 1. Specifically, for the first point, we can take $\boldsymbol{\Gamma}=\mathbf{I}$, which allows the agent to use either the Laplace's method or a covering argument to build an ℓ_2-bonus. For the second point, even if an ℓ_2-bonus can still be possible, the resulting bonus will be less tight than the ℓ_1 one. Thus, we can see a difference (by a factor of $m\log^{-2}(m)$, as mentioned before) between these two classes in terms of regret upper bound that can be reached by the agent. We will see here that this difference is inherent to the classes, and that the two approaches (ℓ_1 and ℓ_2) are matching the respective lower bounds of the classes.

In Table 5.1, we show two existing gap-dependent lower bounds on R_T that depend on the respective class. They are valid for at least one combinatorial structure $\mathcal{A}\subset$

$\mathcal{P}([n])$ such that $m = \max_{A\in\mathcal{A}}|A|$, one distribution $\mathbb{P}_{\mathbf{X}}$ having all gaps equal to Δ and for any consistent policy (Lai and Robbins, 1985), for which the regret on any problem of the class verifies $R_T = o(T^a)$ as $T \to \infty$ for all $a > 0$. We also show in Table 5.1 the regret upper bound reached by the optimistic approaches, showing that they are essentially tight. In Table 5.1, the first lower bound is proved in Combes et al. (2015), and the second one is proved in Kveton, Wen, Ashkan, and Szepesvari (2015b). Each time, they use an action space $\mathcal{A}$ that is reducing to classical MAB, allowing to use the Lai and Robbins (1985) lower bound (Theorem 1).

Let us note finally that in their paper, Degenne and Perchet (2016b) have unified the two above lower bounds by interpolating between the two regimes thanks to a parameter related to the matrix $\mathbf{\Gamma}$. More precisely, their lower bound is of order $\Omega\left(\frac{\sum_i \Gamma_{ii}\log(T)}{\Delta}\right)(1+\gamma(m-1))$, where we already defined γ after (5.1). Their policy, OLS-UCB, uses the bonus from Laplace's method and has a regret upper bound matching this lower bound (up to a polylogarithmic factor in m): using the upper bound (5.1), we can apply Theorem 12 and 13 to obtain the regret bound

$$\mathcal{O}\left(\frac{\log T}{\Delta}\sum_{i\in[n]}\Gamma_{ii}\Big((1-\gamma)\log^2(m)+\gamma m\Big)\right). \tag{5.4}$$

5.3.2 Submodular maximization

In this subsection, we discuss the efficiency of existing algorithms matching the lower bounds in Table 5.1. We consider that an algorithm is *efficient* as soon as the time and space complexity for each round t is polynomial in n and polylogarithmic[2] in t. Notice that the per-round complexity depends substantially on $\mathcal{A}$. We assume $\mathcal{A}$ is such that linear optimization problems on $\mathcal{A}$ — of the form $\max_{A\in\mathcal{A}}\mathbf{e}_A^\mathsf{T}\boldsymbol{\xi}$ for some $\boldsymbol{\xi}\in\mathbb{R}^n$ — can be solved efficiently. As a consequence, an agent knowing $\boldsymbol{\mu}^*$ can efficiently compute A^*. Assuming efficient linear maximization is crucial (cf. Neu and Bartók, 2013; Combes et al., 2015; Kveton, Wen, Ashkan, and Szepesvari, 2015b; Degenne and Perchet, 2016b). Without this assumption, e.g., for $\mathcal{A}$ being dominating sets in a graph, even the offline problem cannot be solved efficiently, and we would have to consider the notion of approximation regret instead, as was done by Chen, Wang, and Yuan (2013).

Many combinatorial semi-bandit algorithms can be seen as a special case of Algorithm 5 for different confidence regions $\mathcal{C}_t$ around $\overline{\boldsymbol{\mu}}_{t-1}$.

Algorithm 5 Generic confidence-region-based algorithm.

At each round t :

Find a confidence region $\mathcal{C}_t \subset \mathbb{R}^n$.

Solve the bilinear program

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

Play A_t.

In particular, CUCB (class $(ii)+(v)$), CUCB-V and CUCB-KL (class $(ii)+(iii)$) are such examples where $\mathcal{C}_t$ is a hypercube. ESCB-KL (Combes et al., 2015) (class

[2] In streaming settings with near real-time requirements, it is imperative to have algorithms that can run with a complexity that stay almost constant across rounds.

$(i)+(iii)$) is defined by the following kl-ball (that resembles to an ellipsoid):

$$\mathcal{C}_t \triangleq \left\{ \boldsymbol{\xi} \in [0,1]^n, \ \sum_{i\in[n]} \mathrm{kl}\Big(\overline{\mu}_{i,t-1}, \xi_i\Big) \leq \delta(t) \right\},$$

where $\delta(t) = \log(t) + 4m\log\log(t)$. We saw that ESCB and OLS-UCB are defined by Algorithm 5 where $\mathcal{C}_t$ is an ellipsoid.

In what follows, we will focus on regions that are axis aligned. In particular, all of the policies mentioned above have such a region, except OLS-UCB when $\boldsymbol{\Gamma}$ is non-diagonal (we can still use the upper bound (5.1) to get an axis aligned region). More precisely, we assume that $\mathcal{C}_t$ is defined through some parameters $p, r \in \{1, \infty\}$, and some functions $g_{i,t}$, $i \in [n]$ by

$$\mathcal{C}_t \triangleq [-r, r]^n \cap \Big(\overline{\boldsymbol{\mu}}_{t-1} + \Big\{\boldsymbol{\xi} \in \mathbb{R}^n, \big\|(g_{i,t}(\xi_i))_i\big\|_p \leq 1\Big\}\Big),$$

where $g_{i,t} = 0$ if $N_{i,t-1} = 0$ and, otherwise, is convex, strictly decreasing on $[-r - \overline{\mu}_{i,t-1}, 0]$ and strictly increasing on $[0, r - \overline{\mu}_{i,t-1}]$ such that $g_{i,t}(0) = 0$. Typically, $r = 1$ under assumption (iii) and $r = \infty$ otherwise.

In Algorithm 5, only A_t needs to be computed. It is a maximizer over $\mathcal{A}$ of the set function

$$\begin{array}{ccc} \mathcal{P}([n]) & \to & \mathbb{R} \\ A & \mapsto & \max_{\boldsymbol{\mu}\in\mathcal{C}_t} \mathbf{e}_A^\mathsf{T}\boldsymbol{\mu} \ . \end{array} \tag{5.5}$$

We can easily evaluate the function (5.5) above for some set $A \in \mathcal{P}([n])$, since it only requires solving a linear optimization problem on the convex[3] set $\mathcal{C}_t$. In Proposition 15, we show that in some cases, the evaluation can be even simpler. However, maximizing the function (5.5) over a combinatorial set $\mathcal{A}$ is not straightforward. Yet, we can already say that when $p = \infty$, this function is linear in A, and therefore can be maximized efficiently. This attests to the efficiency of the CUCB-type approaches. When $p = 1$, we have that the maximization problem is NP-Hard in general (Atamtürk and Gómez, 2017). We will nevertheless extract some interesting properties from this function, in order to be able to maximize it in particular cases. Before studying this function more closely, Definition 15 recalls some well-known properties that can be satisfied by a set function $F : \mathcal{P}([n]) \to \mathbb{R}$.

Definition 15. *A set function F is:*

- *normalized, if $F(\emptyset) = 0$,*
- *linear (or modular) if $F(A) = \mathbf{e}_A^\mathsf{T}\boldsymbol{\xi} + b$, for some $\boldsymbol{\xi} \in \mathbb{R}^n$, $b \in \mathbb{R}$,*
- *non-decreasing if $F(A) \leq F(B)$ $\forall$ $A \subset B \subset [n]$,*
- *submodular if for all $A, B \subset [n]$,*

$$F(A \cup B) + F(A \cap B) \leq F(A) + F(B).$$

Equivalently, F is submodular if for all $A \subset B \subset [n]$, and $i \notin B$, $F(A \cup \{i\}) - F(A) \geq F(B \cup \{i\}) - F(B)$.

[3] $\mathcal{C}_t$ is convex since functions $g_{i,t}$ are convex.

The function (5.5) is clearly normalized, and it can be decomposed into two set functions in the following way,

$$\forall A \subset [n],\ \max_{\boldsymbol{\mu} \in \mathcal{C}_t} \mathbf{e}_A^\intercal \boldsymbol{\mu} = \mathbf{e}_A^\intercal \overline{\boldsymbol{\mu}}_{t-1} + \max_{\boldsymbol{\xi} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}.$$

The linear part $A \mapsto \mathbf{e}_A^\intercal \overline{\boldsymbol{\mu}}_{t-1}$ is efficiently maximized alone, we thus focus on the other part, $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ (the *exploration bonus*). It aims to compensate for the negative selection bias of the first term. We define

$$\begin{aligned}\mathcal{C}_t^+ &\triangleq [-r, r]^n \cap \left(\overline{\boldsymbol{\mu}}_{t-1} + \left\{\boldsymbol{\xi} \in \mathbb{R}_+^n, \left\|(g_{i,t}(\xi_i))_i\right\|_p \leq 1\right\}\right) \\ &= \mathcal{C}_t \cap \left\{\boldsymbol{\mu} \in \mathbb{R}^n,\ \boldsymbol{\mu} \geq \overline{\boldsymbol{\mu}}_{t-1}\right\}\end{aligned}$$

and rewrite $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ through Lemma 8.

Lemma 8. *For all $A \in \mathcal{P}([n])$, $\max_{\boldsymbol{\xi} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi} = \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$.*

The lemma holds as $\left\{(0 \vee \xi_i)_i,\ \boldsymbol{\xi} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}\right\} \subset \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}$. As a corollary, this set function is non-negative, and non-decreasing. It can be written in closed form under additional assumptions, see Proposition 15 and Example 5.

Proposition 15. *Let $A \in \mathcal{P}([n])$, $t \in \mathbb{N}^*$, $p = 1$. Assume that for all $i \in A$, $g_{i,t}$ has a strictly increasing, continuous derivative $g'_{i,t}$ defined on $[0, r - \overline{\mu}_{i,t-1}]$. For $i \in A$, let*

$$f_i(\lambda) \triangleq \begin{cases} g_{i,t}'^{-1}(1/\lambda) & \text{if } 1/\lambda < g'_{i,t}(r - \overline{\mu}_{i,t-1}), \\ r - \overline{\mu}_{i,t-1} & \text{otherwise,} \end{cases}$$

defined for $\lambda \geq 0$. Then, the smallest λ^* satisfying

$$\mathbf{e}_A^\intercal (g_{i,t}(f_i(\lambda^*)))_i \leq 1$$

is such that

$$(\xi_i^*)_i \triangleq (\mathbb{I}\{i \in A\} f_i(\lambda^*))_i \in \operatorname*{arg\,max}_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}.$$

Proof. It suffices to maximize on the coordinates of $\boldsymbol{\xi}$ belonging to A (the others being zero). For all $i \in A$, we let

$$\begin{aligned}\eta_i^* &\triangleq \left(1 - \lambda^* g'_{i,t}(r - \overline{\mu}_{i,t-1})\right) \mathbb{I}\left\{\xi_i^* = r - \overline{\mu}_{i,t-1}\right\} \\ \gamma_i^* &\triangleq \left(\lambda^* g'_{i,t}(0) - 1\right) \mathbb{I}\{\xi_i^* = 0\} = -\mathbb{I}\{\xi_i^* = 0\}.\end{aligned}$$

For all $i \in A$, the function f_i is continuous, non-increasing on $\mathbb{R}_+$, hence so is $\lambda \mapsto \mathbf{e}_A^\intercal (g_{i,t}(f_i(\lambda)))_i$. If $\mathbf{e}_A^\intercal (g_{i,t}(f_i(\lambda^*)))_i < 1$, then necessarily $\lambda^* = 0$. Thus, the following KKT conditions are satisfied:

$$\begin{aligned}\lambda^* \left(\sum_{i \in A} g_{i,t}(\xi_i^*) - 1\right) &= 0, \text{ and} \\ \forall i \in A,\ \lambda^* g'_{i,t}(\xi_i^*) + \eta_i^* - \gamma_i^* &= 1, \\ \eta_i^* (\xi_i^* - r + \overline{\mu}_{i,t-1}) &= 0, \\ -\gamma_i^* \xi_i^* &= 0,\end{aligned}$$

which concludes the proof by the convexity of the constraints and the objective function. □

An important use-case example of Proposition 15 is the following

Example 5. *Let $A \in \mathcal{P}([n])$, $t \in \mathbb{N}^*$. If for all $i \in [n]$, $g_{i,t} = (\cdot)^2 \alpha_{i,t}$ for some $\alpha_{i,t} > 0$, and $r = \infty, p = 1$, then*

$$\max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi} = \sqrt{\mathbf{e}_A^\intercal \left(\frac{1}{\alpha_{i,t}} \right)_i}.$$

Indeed, since the maximizer $\boldsymbol{\xi}^*$ lies at the boundary,

$$\max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi} = \max_{\boldsymbol{\xi} \in \mathbb{R}_+^n,\ \sum_i \alpha_{i,t} \xi_i^2 = 1} \mathbf{e}_A^\intercal \boldsymbol{\xi},$$

and from the first-order optimality condition we deduce that $\mathbf{e}_A = 2\lambda^* (\alpha_{i,t} \xi_i^*)_i$, i.e., $\xi_i^* = \mathbb{I}\{i \in A\}/2\lambda^* \alpha_{i,t}$, where λ^* is necessarily $\frac{1}{2}\sqrt{\mathbf{e}_A^\intercal (1/\alpha_{i,t})_i}$. We thus recover the ESCB*'s exploration bonus for $\alpha_{i,t} = N_{i,t-1}/\delta(t)$.*

Remark 11. *The proof of Proposition 15 follows the same technique as the proof of Theorem 4 by Combes et al. (2015) for developing the computation of the* ESCB-KL *exploration bonus.*

Example 5 is a specific case where the exploration bonus $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ has a particularly simple form: It is the square root of a non-decreasing linear set function. Such a set function is known to be submodular (Stobbe and Krause, 2010). This interesting property helps for maximizing the function (5.5). In Theorem 23, we prove that $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ is in fact always submodular.

Theorem 23. *$A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ is submodular.*

Proof. Let $t \in \mathbb{N}^*$. We consider here the restriction of $g_{i,t}$ to $[0, r - \overline{\mu}_{i,t-1}]$, that we still denote as $g_{i,t}$. Notice that for all $i \in [n]$, $g_{i,t}$ is either 0 or a bijection on $[0, r - \overline{\mu}_{i,t-1}]$ by assumption. For $p = \infty$, we have that

$$\max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi} = \mathbf{e}_A^\intercal \Big(g_{i,t}^{-1}(1) \wedge \Big(r - \overline{\mu}_{i,t-1} \Big) \mathbb{I}\{N_{i,t-1} \geq 1\} + r \mathbb{I}\{N_{i,t-1} = 0\} \Big)_i$$

is a linear set function of A. Assume now that $p = 1$. To show the submodularity of $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ in this case, we will use the notion of *polymatroid*.

Definition 16 (Polymatroid)**.** *A polymatroid is a polytope of the forme*

$$\{\boldsymbol{\xi}' \in \mathbb{R}_+^n, \mathbf{e}_A^\intercal \boldsymbol{\xi}' \leq F(A),\ \forall A \subset [n]\},$$

where F is a non-decreasing submodular function.

Fact 4 (Theorem 3 of He, Zhang, and Zhang, 2012)**.** *Let P be a polymatroid, and let $h_1, \ldots, h_n$ be concave functions. Then $A \mapsto \max_{\boldsymbol{\xi}' \in P} \mathbf{e}_A^\intercal (h_i(\xi_i'))_i$ is submodular.*

Notice that $g_{i,t}^{-1}(\{0\}) = [0, r - \overline{\mu}_{i,t-1}]$ when $N_{i,t-1} = 0$, and that $g_{i,t}^{-1}(\cdot)$ is a strictly increasing concave function on $[0, g_{i,t}(r - \overline{\mu}_{i,t-1})]$, as the inverse function of a strictly increasing convex function when $N_{i,t-1} \geq 1$. So we can rewrite $\mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}$ as an union of product sets:

$$\mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1} = \left\{ \boldsymbol{\xi} \in \prod_{i \in [n]} [0, r - \overline{\mu}_{i,t-1}],\ \sum_{i \in [n]} g_{i,t}(\xi_i) \leq 1 \right\}$$

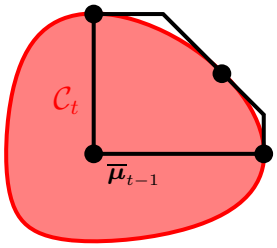

FIGURE 5.1: Illustration of Theorem 23: The confidence region $\mathcal{C}_t$ and the polymatroid defined by the submodular function $A \mapsto \max_{\boldsymbol{\mu} \in \mathcal{C}_t - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\mu}$.

$$= \bigcup_{\substack{\boldsymbol{\xi}' \in \prod_{i\in[n]} [0, g_{i,t}(r-\overline{\mu}_{i,t-1})], \\ \sum_{i\in[n]} \xi'_i \leq 1}} \prod_{i\in[n]} g_{i,t}^{-1}(\{\xi'_i\}).$$

We can thus rewrite our function as

$$\max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi} = \max_{\substack{\boldsymbol{\xi}' \in \prod_{i\in[n]} [0, g_{i,t}(r-\overline{\mu}_{i,t-1})], \\ \sum_{i\in[n]} \xi'_i \leq 1}} \mathbf{e}_A^\intercal \Big(g_{i,t}^{-1}(\xi'_i) \Big)_i,$$

with the convention $g_{i,t}^{-1}(0) = r - \overline{\mu}_{i,t-1}$ when $N_{i,t-1} = 0$.

The constraints' set $\Big\{ \boldsymbol{\xi}' \in \prod_{i\in[n]} [0, g_{i,t}(r - \overline{\mu}_{i,t-1})],\ \sum_{i\in[n]} \xi'_i \leq 1 \Big\}$ is equal to the intersection between $\prod_{i\in[n]} [0, g_{i,t}(r - \overline{\mu}_{i,t-1})]$ and the polymatroid

$$\{ \boldsymbol{\xi}' \in \mathbb{R}_+^n, \mathbf{e}_A^\intercal \boldsymbol{\xi}' \leq \mathbb{I}\{A \neq \emptyset\},\ \forall A \subset [n] \}.$$

This intersection is itself equal to the polymatroid

$$\Big\{ \boldsymbol{\xi}' \in \mathbb{R}_+^n, \mathbf{e}_A^\intercal \boldsymbol{\xi}' \leq \min_{B \subset A} \Big\{ \mathbb{I}\{B \neq A\} + \mathbf{e}_B^\intercal \Big(g_{i,t}(r - \overline{\mu}_{i,t-1}) \Big)_i \Big\},\ \forall A \subset [n] \Big\}.$$

Thus, $\max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$ is the optimal objective value on a polymatroid of a separable concave function, as a function of the index set A. Now, using Fact 4, it is submodular. □

Theorem 23 yields that when the outcome class is strengthen to target the tighter lower bound $n \log(T)/\Delta$, Algorithm 5 reduces to maximizing a submodular set function over $\mathcal{A}$ (the sum of a linear and a submodular function is submodular). Submodular maximization problems have been applied in machine learning before (see e.g., Krause and Golovin, 2011; Bach, 2011), however, maximizing a submodular function F, even for $\mathcal{A} = \{A \in \mathcal{P}([n]), |A| \leq m\}$ and F non-decreasing, is NP-Hard in general (Schrijver, 2008), with an approximation factor of $1 - 1/e$ by the GREEDY algorithm (Nemhauser, Wolsey, and Fisher, 1978).

In the next subsection, with a stronger assumption on $\mathcal{A}$ (but for which submodular maximization is still NP-Hard), we show that both parts of the objective can have different approximation factors. More precisely, we show how to approximate the linear part with factor 1, and the submodular part with a constant factor. This is enough for our purpose, since it simply means that the event $\{\Delta_t \leq 2\text{bonus}_t(A_t)\}$ is replaced by $\{\Delta_t \leq c \cdot \text{bonus}_t(A_t)\}$, where c is a constant.

5.3.3 Efficient algorithms for matroid constraints

In this subsection, we will consider additional structure on $\mathcal{A}$, using the notion of matroid, recalled below.

Definition 17. *A matroid is a pair $([n], \mathcal{I})$, where $\mathcal{I}$ is a family of subsets of $[n]$, called the independent sets, with the following properties:*

- *The empty set is independent, i.e., $\emptyset \in \mathcal{I}$.*
- *Every subset of an independent set is independent, i.e., for all $A \in \mathcal{I}$, if $A' \subset A$, then $A' \in \mathcal{I}$.*
- *If A and B are two independent sets, and $|A| > |B|$, then there exists $x \in A \backslash B$ such that $B \cup \{x\} \in \mathcal{I}$.*

Matroids generalize the notion of linear independence. A maximal (for the inclusion) independent set is called *basis* and all bases have the same cardinality m, which is called the *rank* of the matroid (Whitney, 1935). Many combinatorial problems such as building a spanning tree for network routing (Oliveira and Pardalos, 2005) can be expressed as a linear optimization on a matroid (see Edmonds and Fulkerson, 1965 or Perfect, 1968, for other examples). Let $\mathcal{I} \in \mathcal{P}([n])$ be such that $([n], \mathcal{I})$ forms a matroid. Let $\mathcal{B} \subset \mathcal{I}$ be the set of bases of the matroid $([n], \mathcal{I})$. In the following, we may assume that $\mathcal{A}$ is either $\mathcal{I}$ or $\mathcal{B}$. We also assume that an independence oracle is available, i.e., given an input $A \subset [n]$, it returns TRUE if $A \in \mathcal{I}$ and FALSE otherwise. Maximizing a linear set function L on $\mathcal{A}$ is efficient, and it can be done as follows (Edmonds, 1971): Let σ be the permutation of $[n]$ and j the integer such that $j = m$ in case $\mathcal{A} = \mathcal{B}$ and otherwise, j satisfies

$$\ell_1 \geq \cdots \geq \ell_j \geq 0 \geq \ell_{j+1} \geq \cdots \geq \ell_n,$$

where $\ell_i = L(\{\sigma(i)\})\ \forall i \in [n]$. The optimal S is built greedily starting from $S = \emptyset$, and for $i \in [j]$, $\sigma(i)$ is added to S only if $S \cup \{\sigma(i)\} \in \mathcal{I}$.

Matroid bandits with $\mathcal{A} = \mathcal{B}$ has been studied by Kveton, Wen, Ashkan, Eydgahi, et al. (2014) and Talebi and Proutiere (2016). In this case, the two lower bounds in Table 5.1 coincide to $\Omega(n \log(T)/\Delta)$ (with Δ being the minimal positive gap) and CUCB reaches it. Yet, the efficient implementation of ESCB-style policies are still interesting in this context, for at least two reasons:

- ESCB is, as we'll see, better in practice than CUCB.
- It may provide ideas for efficient implementation of ESCB for other action spaces where CUCB is not known to reach the lower bound.

In the rest of this subsection, we provide efficient approximation routines to maximize the function (5.5) on $\mathcal{A} = \mathcal{I}$ and $\mathcal{B}$ within a factor 1 for the linear part, and a constant factor for the bonus. Therefore, using these routines in Algorithm 5 do not alter its regret upper bound rate.

Let L be a normalized, linear set function, that will correspond to the linear part $A \mapsto \mathbf{e}_A^\mathsf{T} \overline{\boldsymbol{\mu}}_{t-1}$; and let F denote a normalized, non-decreasing, submodular function, that will correspond to the submodular part $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\mathsf{T} \boldsymbol{\xi}$. Unless stated otherwise, we further assume that F is positive (for $A \neq \emptyset$). This is a mild assumption as it holds for $A \mapsto \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\mathsf{T} \boldsymbol{\xi}$ in the unbounded case, i.e., if (iii) is not assumed and $r = \infty$. If (iii) is true, then adding an extra $\mathbf{e}_A^\mathsf{T} \left(\frac{1}{N_i t^{-1^2}} \right)_i$ term will

Algorithm 6 LocalSearch for maximizing $L+F$ on $\mathcal{I}$.

Input: L, F, $\mathcal{I}$, m, $\varepsilon > 0$.
Initialization: $S_{\text{init}} \in \arg\max_{A\in\mathcal{I}} L(A)$.
if $S_{\text{init}} = \emptyset$ **then**
  if $\exists\{x\} \in \mathcal{I}$ such that $(L+F)(\{x\}) > 0$ **then**
    $S_0 \in \arg\max_{\{x\}\in\mathcal{I},\ (L+F)(\{x\})>0} L(\{x\})$.
  else
    Output $\emptyset$
  end if
else
  $S_0 \leftarrow S_{\text{init}}$
end if
$S \leftarrow S_0$.
Repeatedly perform one of the following local improvements **while** possible:

- **Delete an element:**
  if $\exists x \in S$ such that
  $(L+F)(S\backslash\{x\}) > (L+F)(S) + \frac{\varepsilon}{m}F(S)$,
  then $S \leftarrow S\backslash\{x\}$.
  end if
- **Add an element:**
  if $\exists y \in [n]\backslash S,\ S\cup\{y\} \in \mathcal{I}$, such that
  $(L+F)(S\cup\{y\}) > (L+F)(S) + \frac{\varepsilon}{m}F(S)$,
  then $S \leftarrow S\cup\{y\}$.
  end if
- **Swap a pair of elements:**
  if $\exists(x,y) \in S\times[n]\backslash S,\ S\backslash\{x\}\cup\{y\} \in \mathcal{I}$, such that $(L+F)(S\backslash\{x\}\cup\{y\}) > (L+F)(S) + \frac{\varepsilon}{m}F(S)$ **then** $S \leftarrow S\backslash\{x\}\cup\{y\}$ **end if**

end while
Output: S.

recover positivity and increase the regret upper bound by only an additive constant. In the following subsections, we will provide algorithms that efficiently outputs S such that

$$L(S) + \kappa F(S) \geq L(O) + F(O), \quad \forall O \in \mathcal{A}, \tag{5.6}$$

with some appropriate approximation factor $\kappa \geq 1$. It is possible to efficiently output S_1 and S_2 such that we get $L(S_1) \geq L(O_1)$ and $\kappa F(S_2) \geq F(O_2)$ for any $O_1, O_2 \in \mathcal{A}$. Although we can take $O_1 = O_2$, S_1 and S_2 are not necessarily equal, and (5.6) is not straightforward.

Local Search Algorithm

Here, we assume that $\mathcal{A} = \mathcal{I}$. In Algorithm 6, we provide a specific instance of LocalSearch that we tailored to our needs to approximately maximize $L+F$. It starts from the greedy solution $S_{\text{init}} \in \arg\max_{A\in\mathcal{I}} L(A)$. Then, Algorithm 6 repeatedly tries three basic operations in order to improve the current solution. Since every $S \in \mathcal{A}$ can potentially be visited, only *significative* improvements are considered, i.e., improvements greater than $\frac{\varepsilon}{m}F(S)$ for some input parameter $\varepsilon > 0$. The smaller ε is, the higher complexity will be. Notice the improvement threshold $\frac{\varepsilon}{m}F(S)$ does not

depend on L. In fact, this is crucial to ensure that the approximation factor of L is 1. However, this can increase the time complexity. To prevent this increase, the second essential ingredient is the initialization, where only L plays a role. In Theorem 24, we state the approximation guarantees for Algorithm 6 and its time complexity. For $\mathcal{C}_t$ given by any algorithm previously considered, $F(A) = \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\mathsf{T} \boldsymbol{\xi}$, and $\varepsilon = 1$, the time complexity is bounded by $\mathcal{O}(m^2 n \log(mt))$, and the algorithm is thus efficient. Theorem 24 gives a parameter κ arbitrary close to 2 in (5.6).[4]

Theorem 24. *Algorithm 6 outputs $S \in \mathcal{I}$ such that*

$$L(S) + 2(1+\varepsilon)F(S) \geq L(O) + F(O), \quad \forall O \in \mathcal{I}.$$

Its complexity is $\mathcal{O}\Big(mn \log\Big(\frac{\max_{A \in \mathcal{I}} F(A)}{F(S_0)}\Big) / \log(1 + \frac{\varepsilon}{m})\Big)$.

Before proving Theorem 24, we state some well known results about submodular optimization on a matroid.

Proposition 16. *Let $A, B \subset [n]$. If F is submodular, then*

$$\sum_{b \in B \backslash A} (F(B) - F(B \backslash \{b\})) \leq F(B) - F(A \cap B),$$
$$\sum_{a \in A \backslash B} (F(B \cup \{a\}) - F(B)) \geq F(A \cup B) - F(B).$$

Proof. Let $(b_1, \ldots, b_{|B \backslash A|})$ be an ordering of $B \backslash A$. Then, by submodularity of F,

$$\begin{aligned} \sum_{i=1}^{|B \backslash A|} (F(B) - F(B \backslash \{b_i\})) &\leq \sum_{i=1}^{|B \backslash A|} (F(B \backslash \{b_1, \ldots, b_{i-1}\}) - F(B \backslash \{b_1, \ldots, b_i\})) \\ &= F(B) - F(A \cap B). \end{aligned}$$

In the same way, let $(a_1, \ldots, a_{|A \backslash B|})$ be an ordering of $A \backslash B$. Then, by submodularity of F,

$$\begin{aligned} \sum_{i=1}^{|A \backslash B|} (F(B \cup \{a_i\}) - F(B)) &\geq \sum_{i=1}^{|A \backslash B|} (F(B \cup \{a_1, \ldots, a_i\}) - F(B \cup \{a_1, \ldots, a_{i-1}\})) \\ &= F(A \cup B) - F(B). \end{aligned}$$

□

Fact 5 (Theorem 1 of Lee et al., 2010)**.** *Let $A, B \in \mathcal{A}$. Then, there exists a mapping $\alpha : B \backslash A \to A \backslash B \cup \{\emptyset\}$ such that*

- $\forall b \in B \backslash A,\ A \backslash \{\alpha(b)\} \cup b \in \mathcal{A}$
- $\forall a \in A \backslash B,\ |\alpha^{-1}(a)| \leq 1$.

[4] We could design a different version of Algorithm 6, based on Non-ObliviousLocalSearch (Filmus and Ward, 2012) , in order to get κ arbitrary close to $1 + 1/(e-1)$, but with a much worst time complexity. Actually, Sviridenko, Vondrák, and Ward (2013) proposed such an approach, with an approximation factor for L arbitrary close to 1, but not equal, so we would get back the undesirable term, which would require a complexity polynomial in t to control.

Proposition 17. *Let $A, B \in \mathcal{A}$. Let F be a submodular function and $\alpha : B\backslash A \to A\backslash B \cup \{\emptyset\}$ be the mapping given in Fact 5. Then,*

$$\sum_{b\in B\backslash A}(F(A) - F(A\backslash\{\alpha(b)\}\cup\{b\})) + \sum_{a\in A\backslash B,\ \alpha^{-1}(a)=\emptyset}(F(A) - F(A\backslash\{a\}))$$
$$\leq 2F(A) - F(A\cup B) - F(A\cap B).$$

Proof. We decompose $\sum_{b\in B\backslash A}(F(A) - F(A\backslash\{\alpha(b)\}\cup\{b\}))$ into sum of two terms,

$$\sum_{b\in B\backslash A}(F(A) - F(A\backslash\{\alpha(b)\})) + \sum_{b\in B\backslash A}(F(A\backslash\{\alpha(b)\}) - F(A\backslash\{\alpha(b)\}\cup\{b\})).$$

Remark that the first part is equal to

$$\sum_{a\in\alpha(B\backslash A)}(F(A) - F(A\backslash\{a\})) = \sum_{a\in A\backslash B,\ \alpha^{-1}(a)\neq\emptyset}(F(A) - F(A\backslash\{a\})).$$

Thus, together with $\sum_{a\in A\backslash B,\ \alpha^{-1}(a)=\emptyset}(F(A) - F(A\backslash\{a\}))$, we get that

$$\sum_{b\in B\backslash A}(F(A) - F(A\backslash\{\alpha(b)\}\cup\{b\})) + \sum_{a\in A\backslash B,\ \alpha^{-1}(a)=\emptyset}(F(A) - F(A\backslash\{a\}))$$

is equal to

$$\sum_{a\in A\backslash B}(F(A) - F(A\backslash\{a\})) + \sum_{b\in B\backslash A}(F(A\backslash\{\alpha(b)\}) - F(A\backslash\{\alpha(b)\}\cup\{b\})).$$

Finally, we upper bound the first term by $F(A) - F(A\cap B)$ using first inequality of Proposition 16, and the second term by $F(A) - F(A\cup B)$ using first, the submodularity of F to remove $\alpha(b)$ in the summands, and then the second inequality of Proposition 16. □

Proof of Theorem 24. The proof is divided into two parts:

Approximation guarantee If Algorithm 6 outputs $\emptyset$ before entering in the while loop, then by submodularity, for any $S \in \mathcal{I}$,

$$(L+F)(S) \leq \sum_{x\in S}(L+F)(\{x\}) \leq 0.$$

Thus, $\emptyset$ is a maximizer of $L+F$.

Otherwise, the output S of Algorithm 6 satisfies the local optimality of the while loop. We apply Proposition 17 with $A = S$ and $B = O$ for L and F separately,

$$\sum_{b\in O\backslash S}(L(S) - L(S\backslash\{\alpha(b)\}\cup\{b\})) + \sum_{a\in S\backslash O,\ \alpha^{-1}(a)=\emptyset}(L(S) - L(S\backslash\{a\}))$$
$$\leq 2L(S) - L(S\cup O) - L(S\cap O),$$

$$\sum_{b\in O\backslash S}(F(S) - F(S\backslash\{\alpha(b)\}\cup\{b\})) + \sum_{a\in S\backslash O,\ \alpha^{-1}(a)=\emptyset}(F(S) - F(S\backslash\{a\}))$$
$$\leq 2F(S) - F(S\cup O) - F(S\cap O).$$

Then, we sum these two inequalities,

$$\begin{aligned}
&\sum_{b\in O\setminus S}((L+F)(S)-(L+F)(S\setminus\{\alpha(b)\}\cup\{b\}))\\
&+\sum_{a\in S\setminus O,\ \alpha^{-1}(a)=\emptyset}((L+F)(S)-(L+F)(S\setminus\{a\}))\\
&\leq 2(L+F)(S)-(L+F)(S\cup O)-(L+F)(S\cap O)\\
&=2F(S)-F(S\cup O)-F(S\cap O)+L(S)-L(O),
\end{aligned}$$

where the last equality uses linearity of L. Since F is increasing and non-negative, $F(S\cup O)+F(S\cap O)\geq F(O)$, and we get

$$\begin{aligned}
&\sum_{b\in O\setminus S}((L+F)(S)-(L+F)(S\setminus\{\alpha(b)\}\cup\{b\}))\\
&+\sum_{a\in S\setminus O,\ \alpha^{-1}(a)=\emptyset}((L+F)(S)-(L+F)(S\setminus\{a\}))\\
&\leq 2F(S)-F(O)+L(S)-L(O).
\end{aligned}$$

From the local optimality of S, the left hand term in this inequality is lower bounded by

$$\sum_{b\in O\setminus S}\frac{-\varepsilon}{m}F(S)+\sum_{a\in S\setminus O,\ \alpha^{-1}(a)=\emptyset}\frac{-\varepsilon}{m}F(S)\geq -2\varepsilon F(S).$$

The last statement finishes the proof for the approximation inequality.

Time complexity Computing S_0 has a negligible complexity compared to the while loop. The following lemma gives a characterization of S_0.

Lemma 9. $S_0\in\arg\max\{L(A),\ A\in\mathcal{I},\ (F+L)(A)>0\}$.

Proof. From Algorithm 6, if $S_{\text{init}}\neq\emptyset$, then $S_0=S_{\text{init}}$ and $L(S_0)=\max_{A\in\mathcal{I}}L(A)\geq 0$. Thus, $F(S_0)>0$ by assumption on F, giving $(F+L)(S_0)>0$, which ends the proof. If $S_{\text{init}}=\emptyset$, then $L(S_0)=\max\{L(\{x\}),\ \{x\}\in\mathcal{I},\ (L+F)(\{x\})>0\}$. Let $A\in\arg\max\{L(A),\ A\in\mathcal{I},\ (F+L)(A)>0\}$. A is clearly non-empty, and by submodularity of $F+L$, there exists $x\in A$ such that $(F+L)(\{x\})>0$. L is non-increasing from $S_{\text{init}}=\emptyset$, so we get $L(\{x\})\geq L(A)$, which means there is a singleton $\{x\}$ in $\arg\max\{L(A),\ A\in\mathcal{I},\ (F+L)(A)>0\}$, so

$$S_0\in\arg\max\{L(A),\ A\in\mathcal{I},\ (F+L)(A)>0\},$$

which finishes the proof. □

From this lemma, necessarily $L(S_0)\geq L(S_\ell)$ for every iterations $\ell\geq 1$, since the sequence $(L(S_\ell)+F(S_\ell))_\ell$ is increasing, and thus $(F+L)(S_\ell)>0,\ \forall\ell\geq 1$. At each iteration $\ell\geq 1$, Algorithm 6 constructs S_ℓ such that

$$F(S_\ell)>\left(1+\frac{\varepsilon}{m}\right)F(S_{\ell-1})+L(S_{\ell-1})-L(S_\ell).$$

Thus, we must have

$$F(S_\ell)-\left(1+\frac{\varepsilon}{m}\right)^\ell F(S_0)$$

Algorithm 7 GREEDY for maximizing $L + F$ on $\mathcal{B}$.

Input: L, F, $\mathcal{I}$, m.
Initialization: $S \leftarrow \emptyset$.
for $i \in [m]$ **do**
  $x \in \arg\max_{x \notin S,\ S \cup \{x\} \in \mathcal{I}} (L + F)(S \cup \{x\})$.
  $S \leftarrow S \cup \{x\}$.
end for
Output: S.

$$
\begin{aligned}
&\geq \sum_{j=1}^{\ell} \left(1 + \frac{\varepsilon}{m}\right)^{\ell - j} (L(S_{j-1}) - L(S_j)) \\
&= L(S_0)\left(1 + \frac{\varepsilon}{m}\right)^{\ell - 1} - \frac{\varepsilon}{m} \sum_{j=1}^{\ell - 1} L(S_j)\left(1 + \frac{\varepsilon}{m}\right)^{\ell - j - 1} - L(S_\ell) \\
&\geq L(S_0)\left(1 + \frac{\varepsilon}{m}\right)^{\ell - 1} - \frac{\varepsilon}{m} \sum_{j=1}^{\ell - 1} L(S_0)\left(1 + \frac{\varepsilon}{m}\right)^{\ell - j - 1} - L(S_0) = 0,
\end{aligned}
$$

where the last inequality uses $L(S_0) \geq L(S_\ell)$, $\forall \ell \geq 1$. This gives the following upper bound on the number of iteration ℓ:

$$
\ell \leq \frac{\log\left(\frac{F(S_\ell)}{F(S_0)}\right)}{\log(1 + \frac{\varepsilon}{m})} \leq \frac{\log\left(\frac{\max_{A \in \mathcal{A}} F(A)}{F(S_0)}\right)}{\log(1 + \frac{\varepsilon}{m})}.
$$

Finally, the result follows remarking that time complexity per iteration is $\mathcal{O}(mn)$. □

Greedy Algorithm

Here, we assume that $\mathcal{A} = \mathcal{B}$. This situation happens, for instance, under a non-negativity assumption on L, i.e., if we consider non-negative outcomes X_i. We show that the standard GREEDY algorithm (Algorithm 7) improves over Algorithm 6 by exploiting this extra constraint, both in terms of the running time and the approximation factor. We state the result in Theorem 25. Notice that another advantage is that we do not need to assume $F(A) > 0$ for $A \neq \emptyset$ here.

Theorem 25. *Algorithm 7 outputs $S \in \mathcal{B}$ such that*

$$
L(S) + 2F(S) \geq L(O) + F(O), \quad \forall O \in \mathcal{B}.
$$

Its complexity is $\mathcal{O}(mn)$.

As we did previously, before starting the proof of Theorem 25, we state some useful results.

Fact 6 (Brualdi's lemma)**.** *Let $A, B \in \mathcal{B}$. Then, there exists a bijection $\beta : A \to B$ such that*

$$
\forall a \in A,\ A \backslash \{a\} \cup \{\beta(a)\} \in \mathcal{B}.
$$

Furthermore, β is the identity on $A \cap B$.

Proof. A proof is given by Brualdi (1969) and is also proved by Schrijver (2008), as Corollary 39.12a. □

Proposition 18. *Let $A, B \in \mathcal{B}$. Let F be a submodular function and $\beta : A \to B$ be the mapping given in Fact 6. Let $a_1, \ldots, a_k$ be elements of A, and $A_i = \{a_1, \ldots, a_i\}$.*

Then,

$$\sum_{i\in[k]}(F(A_i)-F(A_{i-1}\cup\{\beta(a_i)\}))\leq 2F(A)-F(A\cup B)-F(\emptyset).$$

Proof. We can split $\sum_{i\in[k]}(F(A_{i-1}\cup\{a_i\})-F(A_{i-1}\cup\{\beta(a_i)\}))$ into two terms,

$$\sum_{i=1}^{k}(F(A_{i-1}\cup\{a_i\})-F(A_{i-1}))+\sum_{i=1}^{k}(F(A_{i-1})-F(A_{i-1}\cup\{\beta(a_i)\})).$$

The first term is equal to $F(A_k)-F(\emptyset)$. Using submodularity of F, the second term is upper bounded by

$$\sum_{i=1}^{k}(F(A_m)-F(A_m\cup\{\beta(a_i)\})),$$

which is upper bounded by $F(A_k)-F(A_k\cup B)$ thanks to Proposition 16 and its second inequality. □

Proof of Theorem 25. The time complexity proof is trivial. Let $S_i\triangleq\{s_1,\dots,s_i\}$ be the set maintained in Algorithm 7 after i iterations. Instatianting Proposition 18 with $A_i=S_i$ and $B=O$, we have

$$\sum_{i\in[k]}(F(S_i)-F(S_{i-1}\cup\{\beta(s_i)\}))\leq 2F(S)-F(S\cup O)-F(\emptyset). \tag{5.7}$$

Furthermore, we also have, by linearity of L, and bijectivity of β,

$$\sum_{i\in[k]}(L(S_i)-L(S_{i-1}\cup\{\beta(s_i)\}))=\sum_{i\in[k]}(L(\{s_i\})-L(\{\beta(s_i)\}))=L(S)-L(O). \tag{5.8}$$

Thus, we can sum up (5.7) and (5.8) to get

$$\begin{aligned}&\sum_{i\in[k]}((L+F)(S_i)-(L+F)(S_{i-1}\cup\{\beta(s_i)\}))\\&\leq 2F(S)-F(S\cup O)-F(\emptyset)+L(S)-L(O)\\&\leq 2F(S)-F(O)+L(S)-L(O),\end{aligned}$$

where the last inequality uses the fact that F is increasing and $F(\emptyset)=0$. We finish the proof remarking that by definition of Algorithm 7,

$$(L+F)(S_i)-(L+F)(S_{i-1}\cup\{\beta(s_i)\})\geq 0.$$

□

t-free complexity for $\mathcal{I}$-constraint using greedy The GREEDY algorithm can also be used in the case that $\mathcal{A}=\mathcal{I}$. Indeed, see that GREEDY computes an increasing sequence $\emptyset=S_0\subset\cdots\subset S_m=S$. Instead of outputting S_m, we can output S_i that maximizes $L+2F$. This way, we can write

$$L(S_i)+2F(S_i)\geq L(S_{|O|})+2F(S_{|O|})\geq L(O)+F(O),\quad\forall O\in\mathcal{I}.$$

The second inequality is by using the property of GREEDY (Theorem 25) on the independent set $\mathcal{I}_O \triangleq \mathcal{I} \cap \{A \in \mathcal{P}([n]),\ |A| \leq |O|\}$, noticing that both O and $S_{|O|}$ are bases of the matroid $([n], \mathcal{I}_O)$. This leads to a complexity of order nm, improving over the $m^2 n \log(mt)$ that we got previously.

5.4 The budgeted setting

In this section, we extend results of the two previous subsections to budgeted matroid semi-bandits. In budgeted bandits with single resource and infinite horizon (Ding, Qin, et al., 2013; Xia, Ding, et al., 2016), each arm is associated with both a reward and a cost. The agent aims at maximizing the cumulative reward under a budget constraint for the cumulative costs. Xia, Qin, et al. (2016) studied budgeted bandits with multiple play, where an m-subset A of arms is selected at each round. An optimal (up to a constant term) offline algorithm chooses the same action A^* within each round, where A^* is the minimizer of the ratio "expected cost paid choosing A" over "expected reward gained choosing A". In the setting of Xia, Qin, et al. (2016), the agent observes the *partial* random cost and reward of each arm in A (i.e., semi-bandit feedback), pays the sum of partial costs of A and gains the sum of partial rewards of A. A^* can be computed efficiently, and a Xia, Qin, et al. (2016) give an algorithm based on CUCB. It minimizes the ratio where the averages are replaced by UCBs. We extend this setting to matroid constraints. We assume that total costs/rewards are non-negative linear set functions of the chosen action A. The objective is to minimize a ratio of linear set functions. As previously, two kinds of constraint can be considered for the minimization: either $\mathcal{A} = \mathcal{I}$ or $\mathcal{A} = \mathcal{B}$. Theorem 23 implies that an optimistic estimation of this ratio is of the form $\frac{L_1 - F_1}{L_2 + F_2}$, where for $i \in \{1, 2\}$, F_i are positive (except for $\emptyset$), normalized, non-decreasing, submodular; and L_i are non-negative and linear. $L_1 - F_1$ is a high-probability lower bound on the expected cost paid, and $L_2 + F_2$ is a high-probability upper bound on the expected reward gained. Notice that the numerator, $L_1 - F_1$, can be negative, which can be an incitement to take arms with a high cost/low rewards. Therefore, we consider the minimization of the surrogate $0 \vee \left(\frac{L_1 - F_1}{L_2 + F_2}\right)$. Indeed, $(L_1 - F_1)/(L_2 + F_2)$ is a high probability lower bound on the ratio of expectation, so by monotonicity of $x \mapsto 0 \vee x$ on $\mathbb{R}$, $0 \vee (L_1 - F_1)/(L_2 + F_2)$ is also a high-probability lower bound. We assume L_2 is normalized, but not necessarily L_1. $L_1(\emptyset)$ can be seen as an entry price for a round.

Remark 12. *Notice, If $\mathcal{A} = \mathcal{I}$, and L_1 is normalized, then there is an optimal solution of the form $\{s\} \in \mathcal{I}$ (assuming $\emptyset$ is not feasible): If $L_1 - F_1$ is negative for some $S = \{s\} \subset \mathcal{I}$, then such S is a minimizer. Otherwise, by submodularity (and thus subadditivity, since we consider normalized functions), $L_1 - F_1$ is non-negative, and we have*

$$\begin{aligned}\frac{L_1(S) - F_1(S)}{L_2(S) + F_2(S)} &\geq \frac{\sum_{s \in S} L_1(\{s\}) - F_1(\{s\})}{\sum_{s \in S} L_2(\{s\}) + F_2(\{s\})} \\ &\geq \min_{s \in S} \frac{L_1(\{s\}) - F_1(\{s\})}{L_2(\{s\}) + F_2(\{s\})}.\end{aligned}$$

This minimization problem is at least as difficult as previous submodular maximization problems, taking $L_1 = 1$ and $F_1 = 0$. In order to use our approximation algorithms, we consider the *Lagrangian function* associated to the problem (see e.g., Fujishige, 2005),

$$\mathcal{L}(\lambda, S) \triangleq L_1(S) - F_1(S) - \lambda(L_2(S) + F_2(S)),$$

for $\lambda \geq 0$ and $S \subset [n]$. For a fixed $\lambda \geq 0$, $-\mathcal{L}(\lambda, \cdot)$ is a sum of a linear and a submodular function, and both Algorithms 6 and 7 can be used. However, the first step is to find λ sufficiently close to the optimal ratio

$$\lambda^* = \min_{A \in \mathcal{A}} \left(\frac{L_1(A) - F_1(A)}{L_2(A) + F_2(A)} \right) \vee 0.$$

Remark 13. *For some* $\lambda \geq 0$,

$$\min_{A \in \mathcal{A}} \mathcal{L}(\lambda, A) \geq 0 \Rightarrow \lambda \leq \lambda^*,$$

$$\min_{A \in \mathcal{A}} \mathcal{L}(\lambda, A) \leq 0 \Rightarrow \begin{cases} \lambda \geq \lambda^*, \ \textit{or} \\ \min_{A \in \mathcal{A}} L_1(A) - F_1(A) \leq 0, \\ \textit{which further gives } \lambda^* = 0. \end{cases}$$

From Remark 13, if it was possible to compute $\min_{A \in \mathcal{A}} \mathcal{L}(\lambda, A)$ exactly, then a binary search algorithm would find λ^*. This dichotomy method can be extended to $\kappa-$approximation algorithms by defining the *approximation Lagrangian* as

$$\mathcal{L}_\kappa(\lambda, S) \triangleq L_1(S) - \kappa F_1(S) - \lambda(L_2(S) + \kappa F_2(S)),$$

for $\lambda \geq 0$ and $S \subset [n]$. The idea is to use the following approximation guarantee for a $\kappa-$approximation algorithms outputing S (with objective function $-\mathcal{L}$),

$$\min_{A \in \mathcal{A}} \mathcal{L}_\kappa(\lambda, A) \leq \mathcal{L}_\kappa(\lambda, S) \leq \min_{A \in \mathcal{A}} \mathcal{L}(\lambda, A).$$

Thus, for a given λ, either the l.h.s is strictly negative or the r.h.s is non-negative, depending on the sign of $\mathcal{L}_\kappa(\lambda, S)$. Therefore, from Remark 13, a lower bound λ_1 on λ^*, and an upper bound λ_2 on $\min_{A \in \mathcal{A}} 0 \vee \left(\frac{L_1(A) - \kappa F_1(A)}{L_2(A) + \kappa F_2(A)} \right)$ can be computed. The detailed method is given in Algorithm 8. Notice that it takes as input some $\textsc{Algo}_\kappa$, that can be either Algorithm 6 or Algorithm 7, depending on the assumption on the constraint (either $\mathcal{A} = \mathcal{I}$ or $\mathcal{A} = \mathcal{B}$). We denote the output as $\textsc{Algo}_\kappa(L + F)$, for some linear set function L and some submodular set function F, for maximizing the objective $L + F$ on $\mathcal{A}$, so that $S = \textsc{Algo}_\kappa(L + F)$ satisfies $L(S) + \kappa F(S) \geq \max_{A \in \mathcal{A}} L(A) + F(A)$, i.e., $\kappa = 2(1 + \varepsilon)$ if $\textsc{Algo}_\kappa =$ Algorithm 6, $\mathcal{A} = \mathcal{I}$ and $\kappa = 2$ if $\textsc{Algo}_\kappa =$ Algorithm 7, $\mathcal{A} = \mathcal{B}$. In Theorem 26, we state the result for the output of Algorithm 8.

Theorem 26. *Algorithm 8 outputs* A *such that*

$$0 \vee \left(\frac{L_1(A) - (\kappa + \eta) F_1(A)}{L_2(A) + \kappa F_2(A)} \right) \leq \lambda^*,$$

where λ^* *is the minimum of* $0 \vee \left(\frac{L_1 - F_1}{L_2 + F_2} \right)$ *over* $\mathcal{I}$ *if* $\textsc{Algo}_\kappa =$ *Algorithm 6, and over* $\mathcal{B}$ *if* $\textsc{Algo}_\kappa =$ *Algorithm 7. For* $\mathcal{C}_t$ *given by any policy considered before,* $F(A) = \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\mathsf{T} \boldsymbol{\xi}$*, the complexity is of order* $\log(mt/\eta)$ *times the complexity of* $\textsc{Algo}_\kappa$*.*

Proof. Let A be the output of Algorithm 8 and let

$$\mathcal{L}_{\kappa_1, \kappa_2}(\lambda, S) \triangleq L_1(S) - \kappa_1 F_1(S) - \lambda(L_2(S) + \kappa_2 F_2(S)).$$

Algorithm 8 Binary search for minimizing the ratio $0 \vee (L_1 - F_1)/(L_2 + F_2)$.

Input: L_1, L_2, F_1, F_2, $\textsc{Algo}_\kappa$, $\eta > 0$.
$\delta \leftarrow \frac{\eta \min_{\{s\}\in\mathcal{A}} F_1(\{s\})}{L_2(B)+\kappa^2 F_2(B)}$ with $B = \textsc{Algo}_\kappa(L_2 + \kappa F_2)$.
$A \leftarrow A_0 \in \mathcal{A}\backslash\{\emptyset\}$ arbitrary.
if $\mathcal{L}_\kappa(0, A) > 0$ **then**
  $\lambda_1 \leftarrow 0, \quad \lambda_2 \leftarrow \frac{L_1(A)-F_1(A)}{L_2(A)+F_2(A)}$.
  while $\lambda_2 - \lambda_1 \geq \delta$ **do**
    $\lambda \leftarrow \frac{\lambda_1+\lambda_2}{2}$.
    $S \leftarrow \textsc{Algo}_\kappa(-\mathcal{L}(\lambda, \cdot))$.
    if $\mathcal{L}_\kappa(\lambda, S) \geq 0$ **then**
      $\lambda_1 \leftarrow \lambda$.
    else
      $\lambda_2 \leftarrow \lambda$.
      $A \leftarrow S$.
    end if
  end while
end if
Output: A.

Recall that $\mathcal{L}_\kappa = \mathcal{L}_{\kappa,\kappa}$. Algorithm 8 satisfies either $\mathcal{L}_\kappa(0, A_0) \leq 0$ — in which case Theorem 26 is trivial since

$$0 \vee \left(\frac{L_1(A) - (\kappa+\eta)F_1(A)}{L_2(A) + \kappa F_2(A)}\right) = \lambda^* = 0$$

— or $\mathcal{L}_\kappa(0, A_0) > 0$, in which case we have

$$0 > \mathcal{L}_\kappa(\lambda_2, A) \geq \mathcal{L}_{\kappa+\eta,\kappa}(\lambda_2 - \delta, A) \geq \mathcal{L}_{\kappa+\eta,\kappa}(\lambda_1, A) \geq \mathcal{L}_{\kappa+\eta,\kappa}(\lambda^*, A). \tag{5.9}$$

The first inequality is comes from the update of λ_2: Notice that before the while loop, we have

$$\lambda_2 = \frac{L_1(A_0) - F_1(A_0)}{L_2(A_0) + F_2(A_0)} > \frac{L_1(A_0) - \kappa F_1(A_0)}{L_2(A_0) + \kappa F_2(A_0)} > 0,$$

since $F_2(A_0) > 0$, so $0 > \mathcal{L}_\kappa(\lambda_2, A_0)$ multiplying by $L_2(A_0) + F_2(A_0)$ on both sides. Notice that in particular, this inequality gives that $A \neq \emptyset$.

The second inequality follows from

$$\delta = \frac{\eta \min_{\{s\}\in\mathcal{A}} F_1(\{s\})}{L_2(B) + \kappa^2 F_2(B)} \leq \frac{\eta F_1(A)}{L_2(A) + \kappa F_2(A)}$$

since $A \neq \emptyset$ and $L_2(B) + \kappa^2 F_2(B) \geq L_2(A) + \kappa F_2(A)$. Thus, multiplying by $L_2(A) + \kappa F_2(A) > 0$, and adding $L_1(A) - \kappa F_1(A) - \lambda_2(L_2(A) + \kappa F_2(A))$ gives

$$\begin{aligned} &L_1(A) - (\kappa+\eta)F_1(A) - (\lambda_2 - \delta)(L_2(A) + \kappa F_2(A)) \\ &\leq L_1(A) - \kappa F_1(A) - \lambda_2(L_2(A) + \kappa F_2(A)), \end{aligned}$$

i.e., $\mathcal{L}_{\kappa+\eta,\kappa}(\lambda_2 - \delta, A) \leq \mathcal{L}_\kappa(\lambda_2, A)$.

The third inequality uses $\lambda_2 - \lambda_1 \leq \delta$, and the last inequality uses $\lambda_1 \leq \lambda^*$. Indeed, since $\mathcal{L}_\kappa(\lambda_1, S) \geq 0$, the approximation relation given by $\textsc{Algo}_\kappa$,

$$\mathcal{L}_\kappa(\lambda_1, S) \leq \mathcal{L}(\lambda_1, O),$$

where O is the minimizer of $0 \vee \left(\frac{L_1 - F_1}{L_2 + F_2}\right)$ (for the constraints considered by ALGO$_\kappa$), gives $0 \leq \mathcal{L}(\lambda_1, O)$. Thus,

$$\mathcal{L}^+(\lambda_1, O) \triangleq 0 \vee (L_1(O) - F_1(O)) - \lambda_1(L_2(O) + F_2(O)) \geq \mathcal{L}(\lambda_1, O) \geq 0.$$

Finally, since $L_2(O) + F_2(O) > 0$ ($O \neq \emptyset$), we have $\lambda_1 \leq \lambda^*$.

In (5.9), since $A \neq \emptyset$, we have $\frac{L_1(A) - (\kappa+\eta)F_1(A)}{L_2(A) + \kappa F_2(A)} \leq \lambda^*$ and therefore,

$$0 \vee \left(\frac{L_1(A) - (\kappa + \eta)F_1(A)}{L_2(A) + \kappa F_2(A)}\right) \leq \lambda^*.$$

The time complexity for the binary search is $\mathcal{O}(\log(1/\delta)) \leq \mathcal{O}(\log(mt/\eta))$ for $\mathcal{C}_t$ given by any policy previously considered, and $F(A) = \max_{\boldsymbol{\xi} \in \mathcal{C}_t^+ - \overline{\boldsymbol{\mu}}_{t-1}} \mathbf{e}_A^\intercal \boldsymbol{\xi}$. □

5.5 Experiments and discussion

We provide experiments for a *graphic matroid*, on a five nodes complete graph, as did Combes et al. (2015). We thus have $n = 10$, $m = 4$. We consider two experiments. In the first one we use $\mathcal{A} = \mathcal{B}$, $\mu_i^* = 1 + \Delta \mathbb{I}\{i \leq m\}$, for all $i \in [n]$, and in the second, $\mathcal{A} = \mathcal{I}$, where we set $\mu_i^* = \Delta(2\mathbb{I}\{i \leq m - 1\} - 1)$, $\forall i \in [n]$. We take $\Delta = 0.1$, with outcomes drawn from independent unit variance Gaussian distributions. Figure 5.2 illustrates the comparison between CUCB and our implementations of ESCB (Combes et al., 2015) using Algorithm 7 (left) and 6 (right, with $\varepsilon = 0.1$), showing the behavior of the regret vs. time horizon. We observe that although we are approximating the confidence region within a factor at least 2 (and thus force the exploration), our efficient implementation outperforms CUCB. Indeed, we take advantage of the previously inefficient algorithm (that we made efficient) to use outcome independence, which the more conservative CUCB is not able to. The latter algorithm has still a better per round-time complexity of $\mathcal{O}(n \log m)$ and may be more practical on larger instances.

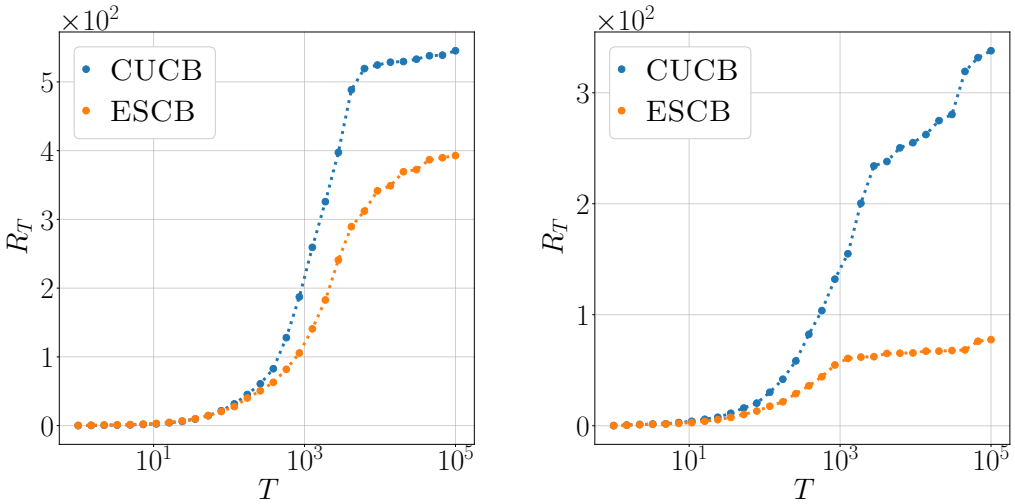

FIGURE 5.2: Cumulative regret for the minimum spanning tree setting in up to 10^5 rounds, averaged over 100 independent simulations. **Left**: for $\mathcal{A} = \mathcal{B}$. **Right**: for $\mathcal{A} = \mathcal{I}$.

5.5.1 Discussion

We gave several approximation schemes for the confidence regions and applied them to combinatorial semi-bandits with matroid constraints and their budgeted version. We believe our approximation methods can be extended to approximation regret for non-linear objective functions (e.g., for influence maximization, Wang and Chen, 2018), if the maximization algorithm keeps the same approximation factor for the objective, either with or without the bonus.

Chapter 6

Budgeted Online Influence Maximization

This chapter aims to further study another well-known example of CMAB-T problem, which is *online influence maximization* (OIM). It is based on our papers Perrault, Healey, et al. (2020a) and Perrault, Healey, et al. (2020b). In particular, we will introduce a new budgeted framework for OIM, considering the total cost of an advertising campaign instead of the common cardinality constraint on a chosen influencer set. Our approach models better the real-world setting where the cost of influencers varies and advertizers want to find the best value for their overall social advertising budget. We propose an algorithm assuming an independent cascade diffusion model and edge-level semi-bandit feedback, and provide both theoretical and experimental results. Our analysis is also valid for the cardinality-constraint setting and improves the state of the art regret bound in this case.

6.1 Problem formulation

Viral marketing through online social networks now represents a significant part of many digital advertising budgets. In this form of marketing, companies incentivize chosen influencers in social networks (e.g., Facebook, Twitter, YouTube) to feature a product in hopes that their followers will adopt the product and repost the recommendation to their own network of followers. The effectiveness of the chosen set of influencers can be measured by the expected number of users that adopt the product due to their initial recommendation, called the *spread*. *Influence maximization* (IM, Kempe, Kleinberg, and Tardos (2015)) is the problem of choosing the optimal set of influencers to maximize the spread under a cardinality constraint on the chosen set.

In order to define the spread, we need to specify a diffusion process such as independent cascade (IC) or linear threshold (LT) (Kempe, Kleinberg, and Tardos, 2015). The parameters of these models are usually *unknown*. Different methods exist to estimate the parameters of the diffusion model from historical data, however historical data is often difficult to obtain. Another possibility is to consider *online influence maximization* (OIM) (Vaswani, Lakshmanan, and Mark Schmidt, 2015; Wen, Kveton, Valko, et al., 2017) where an agent actively learns about the network by interacting with it repeatedly, trying to find the best seed influencers. The agent thus faces the dilemma of exploration versus exploitation, allowing us to see it as *multi-armed bandits* problem (Auer, Cesa-Bianchi, and Fischer, 2002). More precisely, the agent faces IM over T rounds. Each round, it selects m seeds (based on *feedback* from prior rounds) and diffusion occurs; then it gains a reward equal to the spread and receives some feedback on the diffusion.

IM and OIM optimize with the constraint of a fixed number of seeds. This reflects a fixed seed cost model, for example, where influencers are incentivized by being given

an identical free product. In reality, however, many influencers demand different levels of compensation. Those with a high out-degree (e.g., number of followers) are usually more expensive. Due to these cost variations, marketers usually wish to optimize their seed sets S under a budget $c(S) \leq b$ rather than a cardinality constraint $|S| \leq m$. Optimizing a seed set under a budget has been studied in the *offline* case by Nguyen and Zheng (2013). In the *online* case, Wang, Yang, et al. (2020) considered the relaxed constraint $\mathbb{E}[c(S)] \leq b$, where the expectation is over the possible randomness of S.[1] We believe however that the constraint of a fixed, equal budget $c(S) \leq b$ at each round does not sufficiently model the willingness to choose a cost-efficient seed set. Indeed, we see that the choice of b is crucial: a b too large translates into a waste of budget (some seeds that are too expensive will be chosen) and a b too small translates into a waste of time (a whole round is used to influence only a few users). To circumvent this issue, instead of a budget per round, in our framework, we allow the agent to choose seed sets of any cost at each round, under an overall budget constraint (equal to $B = bT$ for instance). In summary, we incorporate the OIM framework into a *budgeted bandit* setting. Our setting is more flexible for the agent, and better meets real-world needs.

6.1.1 Related work on IM

We recall that IM can be formally defined as follows. A social network is modeled as a directed graph $G = (V, E)$, with nodes V representing users and edges E representing connections. An underlying diffusion model D governs how information spreads in G. More precisely, D is a probability distribution on subgraphs G' of G, and given some seed set S, the spread $\sigma(S)$ is defined as the expected number of S-reachable[2] nodes in $G' \sim D$. IM aims to find S that is a solution to

$$\max_{|S|=m} \sigma(S). \tag{6.1}$$

Although IM is NP-hard under standard diffusion models — i.e., IC and LT — σ is a monotone *submodular*[3] function (Fujishige, 2005), and given a value oracle access to σ, the standard GREEDY algorithm solves (6.1) within a $1 - 1/e$ approximation factor (Nemhauser, Wolsey, and Fisher, 1978). There have been multiple lines of work for IM, including the development of heuristics, approximation algorithms, as well as alternative diffusion models (Leskovec, Krause, et al., 2007; Goyal, Lu, and Lakshmanan, 2011; Tang, Xiao, and Shi, 2014; Tang, Shi, and Xiao, 2015). Additionally, there are also results on learning D from data in the case it is not known (Saito, Nakano, and Kimura, 2008; Goyal, Bonchi, and Lakshmanan, 2010; Gomez-Rodriguez, Leskovec, and Krause, 2012; Netrapalli and Sanghavi, 2012).

6.1.2 Related work on OIM

Prior work in OIM has mainly considered either *node level semi-bandit* feedback (Vaswani, Lakshmanan, and Mark Schmidt, 2015), where the agent observes all the S-reachable nodes in G', or *edge level semi-bandit* feedback (Wen, Kveton, Valko, et al., 2017), where the agent observes the whole S-reachable subgraph (i.e., the subgraph of G' induced by S-reachable nodes). Other, weaker, feedback settings

[1] This relaxation is to avoid a computationally costly *partial enumeration* (Krause and Guestrin, 2005; Khuller, Moss, and Naor, 1999)

[2] nodes that are reachable from some node in S.

[3] f is submodular if $f(A \cup \{i\}) - f(A)$ is non-increaing with A.

have also been studied including: pairwise influence feedback, where all nodes that would be influenced by a seed set are observed but not the edges connecting them, i.e., $(\{i\}\text{-reachable nodes})_{i\in S}$ is observed (Vaswani, Kveton, et al., 2017); local feedback, where the agent observes a set of out-neighbors of S (Carpentier and Valko, 2016) and immediate neighbor observation where the agent only observes the out-degree of S (Lugosi, Neu, and Olkhovskaya, 2019).

6.1.3 Our contributions

We define the budgeted OIM paradigm and propose a performance metric for an online policy on this problem using the notion of *approximation regret* (Chen, Wang, and Yuan, 2013). To the best of our knowledge, the both of contributions are new. We then focus our study on the IC model with edge level semi-bandit feedback. We design a CUCB-style algorithm and prove logarithmic regret bounds. We also propose some modifications of this algorithm with improving the regret rates. These gains apply to the non-budgeted setting, giving an improvement over the state-of-the-art analysis of the standard CUCB-approach (Wang and Chen, 2017). Our proof incorporates an approximation guarantee of GREEDY for ratio of submodular and modular functions, which may also be of independent interest.

6.1.4 Problem definition

In this subsection, we formulate the problem of budgeted OIM and give a regret definition for evaluating policies in that setting. We also justify our choice for this notion of regret. We consider a fixed directed network $G = (V, E)$, known to the agent, with $V \triangleq \{1, \dots, |V|\}$. We denote by $ij \in E$ the directed edge from node i to j in G. We assume that G doesn't have self-loops, i.e., for all $ij \in E$, $i \neq j$. For a node $i \in V$, a subset $S \subset V$, and a vector $\mathbf{w} \in \{0,1\}^E$, the predicate $S \overset{\mathbf{w}}{\rightsquigarrow} i$ holds if, in the graph defined by $G_{\mathbf{w}} \triangleq (V, \{ij \in E, w_{ij} = 1\})$, there is a forward path from a node in S to the node i. If it holds, we say that i is influenced by S under $\mathbf{w}$. We define $p_i(S;\mathbf{w}) \triangleq \mathbb{I}\left\{S \overset{\mathbf{w}}{\rightsquigarrow} i\right\}$ and the *spread* as $\sigma(S;\mathbf{w}) \triangleq \left|\left\{i \in V,\ S \overset{\mathbf{w}}{\rightsquigarrow} i\right\}\right|$. Our diffusion process is defined by the random vector $\mathbf{W} \in \{0,1\}^E$, and our cost is defined by the random[4] vector $\mathbf{C} \in [0,1]^{V\cup\{0\}}$ where the added component C_0 represents any fixed costs[5]. Notice, random costs are neither assumed to be mutually independent nor independent from $\mathbf{W}$. We will see that components of $\mathbf{W}$ might however be mutually independent (e.g., for the IC model).

Budgeted online influence maximization

The agent interacts with the diffusion process across several rounds, using a learning policy. At each round $t \geq 1$, the agent first selects a seed set $S_t \subset V$, based on its past observations. Then, the random vectors for both the diffusion process $\mathbf{W}_t \sim \mathbb{P}_{\mathbf{W}}$ and the costs $\mathbf{C}_t \sim \mathbb{P}_{\mathbf{C}}$ are sampled independently from previous rounds. Then, the agent observes some feedback from both the diffusion process and the costs.

We provide in (6.2) the expected cumulative rewards F_B defined for some total budget $B > 0$. The goal for the agent is to follow a learning policy π maximizing F_B. In (6.2), recall that S_t is the seed set selected by π at round t.

[4] Although costs are usually deterministic, we assume randomness for more generality (influencer campaigns may have uncertain surcharges for example).

[5] We provide a toy example where C_0 models a concrete quantity: Assume you want to fill your restaurant. You may pay some seeds and ask them to advertise/influence people. C_0 represents the cost of the food, the staff, the rent, the taxes, ...

$$F_B(\pi) \triangleq \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \sigma(S_t; \mathbf{W}_t)\right]. \tag{6.2}$$

τ_B is the random round at which the remaining budget becomes negative: if $B_t \triangleq B - \sum_{t' \leq t}\left(\mathbf{e}_{S_{t'}}^{\mathsf{T}} \mathbf{C}_{t'} + C_{0,t'}\right)$, then $B_{\tau_B - 1} \geq 0$ and $B_{\tau_B} < 0$. Notice, quantities B_t and τ_B are usual in budgeted multi-armed bandits (Xia, Qin, et al., 2016; Ding, Qiny, et al., 2013).

Performance metric

We restrict ourselves to *efficient* policies, i.e., we consider a complexity constraint on the policy the agent can follow: For a round t, the space and time complexity for computing S_t has to be polynomial in $|V|$, and polylogarithmic in t. To evaluate the performance of a learning policy π, we use the notion of approximation regret (Kakade, Kalai, and Ligett, 2009; Streeter and Golovin, 2009; Chen, Wang, and Yuan, 2016). The agent wants to follow a learning policy π which minimizes

$$R_{B,\varepsilon}(\pi) \triangleq (1 - 1/e - \varepsilon)F_B^* - F_B(\pi),$$

where F_B^* is the best possible value of F_B over all policies (thus leveraging on the knowledge of $\mathbb{P}_{\mathbf{W}}$ and $\mathbb{P}_{\mathbf{C}}$), and where $\varepsilon > 0$ is some parameter the agent can control to determine the tradeoff between computation and accuracy.

Remark 14. *This OIM with a total budget B is different from OIM in previous work, such as Wang and Chen (2017), even when we set all costs to be equal. In our setting, there is only one total budget for all rounds, and the policy is free to choose seed sets of different cost in each round, whereas in the previous work, each round had a fixed budget for the number/cost of seeds selected. Our setting thus avoid the use of a budget per round, which is in practice more difficult to establish than a global budget B. Nevertheless, as we will see in subsection 6.3.2, both types of constraints (global and per round) can be considered simultaneously when the true costs are known.*

Justification for the approximation regret

In the non-budgeted OIM problem with a cardinality constraint given by $m \in [|V|]$, let us recall that the approximation regret

$$R_{T,\varepsilon}(\pi) \triangleq \sum_{t \leq T} \max_{\substack{S \subset V, \\ |S| = m}} \mathbb{E}[(1 - 1/e - \varepsilon)\sigma(S; \mathbf{W}) - \sigma(S_t; \mathbf{W})]$$

is standard (Wen, Kveton, Valko, et al., 2017; Wang and Chen, 2017). In this notion of regret, the factor $(1 - 1/e - \varepsilon)$ (Feige, 1998; Chen, Wang, and Wang, 2010) reflects the difficulty of approximating the following NP-Hard (Kempe, Kleinberg, and Tardos, 2015) problem in the case the distribution $\mathbb{P}_{\mathbf{W}}$ is described by IC or LT, and is known to the agent:

$$\max_{S \subset V,\ |S| = m} \mathbb{E}[\sigma(S; \mathbf{W})]. \tag{6.3}$$

For our budgeted setting, at first sight, it is not straightforward to know which approximation factor to choose. Indeed, since the random horizon may be different in $F_B(\pi)$ and in F_B^*, the expected regret $R_{B,\varepsilon}(\pi)$ is not expressed as the expectation of

a sum of approximation gaps, so we can't directly reduce the regret level approximability to the gap level approximability. We thus consider a quantity provably close to $R_{B,\varepsilon}(\pi)$ and easier to handle.

Proposition 19. *Define*

$$\lambda^* \triangleq \max_{S \subset V} \frac{\mathbb{E}[\sigma(S; \mathbf{W})]}{\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{C}\right]}.$$

For all $S \subset V$, define the gap *corresponding to S as*

$$\Delta(S) \triangleq (1 - 1/e - \varepsilon)\lambda^* \mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{C}\right] - \mathbb{E}[\sigma(S; \mathbf{W})].$$

Then, for any policy π selecting S_t at round t,

$$\left| R_{B,\varepsilon}(\pi) - \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \Delta(S_t)\right] \right| \leq 2|V| + 2\lambda^*(1 + |V|).$$

From Proposition 19 (which is proven in subsection 6.5.1), $R_{B,\varepsilon}(\pi)$ and $\mathbb{E}\left[\sum_{t=1}^{\tau_B-1} \Delta(S_t)\right]$ are equivalent in term of regret upper bound rate. Therefore, the factor $(1 - 1/e - \varepsilon)$ should reflect the approximability of

$$\max_{S \subset V} \mathbb{E}[\sigma(S; \mathbf{W})] / \mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{C}\right] = \max_{S \subset V} f(S)/c(S). \tag{6.4}$$

Considering the specific problem where the cost function is of the form $c(S) = c_1|S| + c_0$, for some $(c_0, c_1) \in [0, 1]^2$, we can reduce[6] the approximability of (6.4) to the approximability of the following problem considered in Wang, Yang, et al. (2020):

$$\max_{S \subset V} \mathbb{E}[f(S)] \ s.t. \ \mathbb{E}[|S|] = m, \tag{6.5}$$

for some given integer m, where the expectations are with respect to a randomization in the approximation algorithm. Wang, Yang, et al. (2020) proved that this problem is NP-hard by reducing to the *set cover* problem. We show here that an approximation ratio α better that $1 - 1/e$ yields a high probability approximation for set cover within $(1 - \delta)\log(|V|)$, $\delta > 0$, which is impossible unless NP $\subset$ BPTIME$\left(n^{\mathcal{O}(\log\log(|V|))}\right)$ (Feige, 1998). Consider the graph where the collection of closed out-neighborhoods is exactly the collection of sets in the set cover instance. First, trying out all possible values of m, we concentrate on the case in which the optimal m for set cover is tried out. As in Feige (1998), for $k \in \mathbb{N}^*$, we repeatedly apply the algorithm that α-approximate (6.5). It outputs a set S_k (that can be associated with a set of neighborhoods) and after each application the nodes already covered by previous applications are removed from the graph, giving a sequence of objective functions (f_k) with $f_1 = f$. We thus obtain

$$\mathbb{E}[f_k(S_k)|S_1, \ldots, S_{k-1}] \geq \alpha\left(|V| - \sum_{k'=1}^{k-1} f_{k'}(S_{k'})\right).$$

[6] We can first easily reduce to one variable, denoted c. The corresponding ratio is denoted $r_c(S)$ and its approximate maximizer is denoted S_c. For a fixed c, with a linear search, we can consider the algorithm outputting $\widetilde{S}_c = \arg\max_{S_{c'}} r_c(S_{c'})$. Then, $c \mapsto \left|\widetilde{S}_c\right|$ is monotone, so for a given m, we can find c s.t. either $\left|\widetilde{S}_c\right| = m$ or $\left|\widetilde{S}_{c^-}\right| < m < \left|\widetilde{S}_{c^+}\right|$. In this last case, we can build a randomized set of expected cardinality m.

Noticing that $\mathbb{E}[f(S_1 \cup \cdots \cup S_k)] = \sum_{k'=1}^{k} \mathbb{E}[f_{k'}(S_{k'})]$, we get

$$\mathbb{E}[f(S_1 \cup \cdots \cup S_k)] \geq \Big(1 - (1-\alpha)^k\Big)|V|.$$

After $\ell = \lfloor \log(1/|V|)/\log(1-\alpha) \rfloor < (1-\delta)\log(|V|)$ iterations, we obtain that $S = S_1 \cup \cdots \cup S_\ell$ is a cover, i.e., $f(S) = |V|$. The result follows noticing that in expectation (and so with probability at least $1/(\ell m)$), we have $|S| \leq \ell m$.

6.2 ℓ_1-based approach

We are now considering algorithms for IC with edge level semi-bandit feedback.

6.2.1 Setting

For $\mathbf{w} \in [0,1]^V$, we recall that we can define an IC model by taking

$$\mathbb{P}_{\mathbf{W}} = \otimes_{ij \in E} \text{Bernoulli}(w_{ij}).$$

We can extend the two previous functions p_i and σ to $\mathbf{w}$ taking values in $[0,1]^V$ as follows: Let $\mathbf{W} \sim \otimes_{ij \in E} \text{Bernoulli}(w_{ij})$. We define the probability that i is influenced by S under $\mathbf{W}$ as $p_i(S;\mathbf{w}) \triangleq \mathbb{P}\Big[S \overset{\mathbf{W}}{\rightsquigarrow} i\Big]$, and we let the spread be $\sigma(S;\mathbf{w}) \triangleq \mathbb{E}\Big[\Big|\Big\{i \in V,\ S \overset{\mathbf{W}}{\rightsquigarrow} i\Big\}\Big|\Big]$. Another expression for the spread is $\sigma(S;\mathbf{w}) = \sum_{i \in V} p_i(S;\mathbf{w})$. We fix a weight vector on E, $\mathbf{w}^* \triangleq \Big(w^*_{ij}\Big)_{ij \in E} \in [0,1]^E$, a cost vector on $V \cup \{0\}$, $\mathbf{c}^* \triangleq (c^*_i)_{i \in V \cup \{0\}} \in [0,1]^{V \cup \{0\}}$, with $c^*_0 > 0$. These quantities are initially *unknown* to the agent. We assume from now that

$$\mathbb{P}_{\mathbf{W}} \triangleq \otimes_{ij \in E} \text{Bernoulli}\Big(w^*_{ij}\Big),$$

and that

$$\mathbb{E}[\mathbf{C}] = \mathbf{c}^*.$$

We also define $S^* \in \arg\max_{S \subset V} \sigma(S;\mathbf{w}^*)/\mathbf{e}_{S \cup \{0\}}^\intercal \mathbf{c}^*$. We assume that the feedback received by the agent at round t is

$$\Big\{W_{ij,t},\ ij \in E,\ S_t \overset{\mathbf{W}_t}{\rightsquigarrow} i\Big\}.$$

The agent also receives semi-bandit feedback from the costs, i.e.,

$$\{C_{i,t},\ i \in V,\ i \in S_t \cup \{0\}\}$$

is observed.

6.2.2 Algorithm design

In this subsection, we present BOIM-CUCB, CUCB for Budgeted OIM problem as Algorithm 9. As we saw in Proposition 19, the policy that, at each round, $(1 - 1/e - \varepsilon)-$approximately maximize

$$S \mapsto \frac{\sigma(S;\mathbf{w}^*)}{\mathbf{e}_S^\intercal \mathbf{c}^* + c^*_0} \tag{6.6}$$

Algorithm 9 BOIM-CUCB

Input: $\varepsilon > 0$, $B_0 = B > 0$.
for each round $t \geq 1$ **do**
If true costs are known, then $\mathbf{c}_t \leftarrow \mathbf{c}^*$.
Compute S_t given by Algorithm 10 with input $S \mapsto \sigma(S; \mathbf{w}_t)$, $\mathbf{c}_t$.
Select seed set S_t, and pay $\mathbf{e}_{S_t \cup \{0\}}^\intercal \mathbf{C}_t$ (i.e., remove this cost from B_{t-1} to get the new budget B_t).
if $B_t \geq 0$, **then**
Get the reward $\sigma(S_t; \mathbf{W}_t)$, get the feedback, and update corresponding quantities accordingly.
else
The budget is exhausted: leave the for loop.
end if
end for

has a bounded regret. Thus, BOIM-CUCB shall be based on this objective. Not only there are some estimation concerns due to the unknown parameters $\mathbf{w}^*, \mathbf{c}^*$, but in addition to that, we also need to evaluate/optimize our estimates of (6.6).

We begin by introducing some notations. We define the empirical means for $t \geq 1$ as: For all $i \in V \cup \{0\}$,

$$\overline{c}_{i,t-1} \triangleq \frac{\sum_{t' \in [t-1]} \mathbb{I}\{i \in S_{t'} \cup \{0\}\} C_{i,t'}}{N_{\ominus i,t-1}},$$

and for all $ij \in E$,

$$\overline{w}_{ij,t-1} \triangleq \frac{\sum_{t' \in [t-1]} \mathbb{I}\left\{S_{t'} \overset{\mathbf{W}_{t'}}{\rightsquigarrow} i\right\} W_{ij,t'}}{N_{\oplus i,t-1}},$$

where $N_{\ominus i,t-1} \triangleq \sum_{t'=1}^{t-1} \mathbb{I}\{i \in S_{t'}\}$, $N_{\oplus i,t-1} \triangleq \sum_{t'=1}^{t-1} \mathbb{I}\left\{S_{t'} \overset{\mathbf{W}_{t'}}{\rightsquigarrow} i\right\}$. Using concentration inequalities, we get confidence intervals for the above estimates. We are then able to use an upper-confidence-bound (UCB) strategy (Auer, Cesa-Bianchi, and Fischer, 2002). More precisely, in the case costs are unknown, we first build the *lower* confidence bound (LCB) on c_i^* as follows

$$c_{i,t} \triangleq 0 \vee \left(\overline{c}_{i,t-1} - \sqrt{\frac{1.5 \log(t)}{N_{\ominus i,t-1}}}\right).$$

We can also define UCBs for w_{ij}^*:

$$w_{ij,t} \triangleq 1 \wedge \left(\overline{w}_{ij,t-1} + \sqrt{\frac{1.5 \log(t)}{N_{\oplus i,t-1}}}\right).$$

We use $w_{ij,t} = 1$ (and $c_{i,t} = 0$) when the corresponding counter is equal to 0. Our BOIM-CUCB approach chooses at each round t the seed set S_t given by Algorithm 10 which, as we shall see, approximately maximize $S \mapsto \sigma(S; \mathbf{w}_t) / \left(\mathbf{e}_{S \cup \{0\}}^\intercal \mathbf{c}_t\right)$. Indeed, with high probability, this set function is an upper bound on the true ratio (6.6) (using that σ is non decreasing w.r.t. $\mathbf{w}$). Notice that this approach is followed by Wang and Chen (2017) for the non budgeted setting, i.e., they choose S_t, $|S_t| \leq m$ that approximately maximize $S \mapsto \sigma(S; \mathbf{w}_t)$. To complete the description of our algorithm, we need to describe Algorithm 10. This is the purpose of the following.

6.2.3 Greedy for ratio maximization

In BOIM-CUCB, one has to approximately maximize the ratio $S \mapsto \sigma(S; \mathbf{w}_t)/\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}_t$, that is a ratio of submodular over modular function. A GREEDY technique can be used (see Algorithm 10). Indeed, instead of maximizing the marginal contribution at each time step, as the standard GREEDY algorithm do, the approach is to maximize the so called bang-per-buck, i.e., the marginal contribution divided by the marginal cost. This builds a sequence of increasing subsets, and the final output is the one that maximizes the ratio. The following Proposition 20 gives an approximation factor of $1 - 1/e$ for Algorithm 10.

Proposition 20. *Algorithm 10 with input $\sigma, \mathbf{c}$ is guaranteed to obtain a solution S such that:*

$$\left(1 - e^{-1}\right) \frac{\sigma(S^*)}{\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}} \leq \frac{\sigma(S)}{\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}}.$$

Proof. In the proof, we use the notation $\sigma(i|S) \triangleq \sigma(\{i\} \cup S) - \sigma(S)$. For any $k \in [|V|]$,

$$\begin{aligned}
\sigma(S^*) - \sigma(S_{k-1}) &\leq \sum_{i \in S^* \setminus S_{k-1}} \sigma(i|S_{k-1}) && \text{Submodularity, monotonicity of } \sigma \\
&\leq \frac{\sigma(i_k|S_{k-1})}{c_{i_k}} \sum_{i \in S^* \setminus S_{k-1}} c_i && \text{Algorithm 10} \\
&\leq \frac{\sigma(i_k|S_{k-1})}{c_{i_k}} \sum_{i \in S^*} c_i.
\end{aligned}$$

i.e., for all $k \in [|V|]$ such that $\sigma(S^*) - \sigma(S_{k-1}) \geq 0$,

$$\frac{c_{i_k}}{\mathbf{e}_{S^*}^\intercal \mathbf{c}} \leq \frac{\sigma(i_k|S_{k-1})}{\sigma(S^*) - \sigma(S_{k-1})}. \tag{6.7}$$

There must be an index $\ell \in \{0, 1, \ldots, |V| - 1\}$ such that $\mathbf{e}_{S_\ell}^\intercal \mathbf{c} \leq \mathbf{e}_{S^*}^\intercal \mathbf{c} \leq \mathbf{e}_{S_{\ell+1}}^\intercal \mathbf{c}$. Let $\beta \in [0, 1]$ be such that

$$\mathbf{e}_{S^*}^\intercal \mathbf{c} = (1 - \beta)\mathbf{e}_{S_\ell}^\intercal \mathbf{c} + \beta \mathbf{e}_{S_{\ell+1}}^\intercal \mathbf{c}. \tag{6.8}$$

If $\sigma(S^*) - (1 - \beta)\sigma(S_\ell) - \beta\sigma(S_{\ell+1}) \leq 0$, then we have

$$\left(1 - e^{-1}\right) \frac{\sigma(S^*)}{\mathbf{e}_{S^*}^\intercal \mathbf{c}} \leq \frac{\sigma(S^*)}{\mathbf{e}_{S^*}^\intercal \mathbf{c}} \leq \frac{(1 - \beta)\sigma(S_\ell) + \beta\sigma(S_{\ell+1})}{(1 - \beta)\mathbf{e}_{S_\ell}^\intercal \mathbf{c} + \beta \mathbf{e}_{S_{\ell+1}}^\intercal \mathbf{c}}.$$

Else,

$$\sigma(S^*) - (1 - \beta)\sigma(S_\ell) - \beta\sigma(S_{\ell+1}) > 0 \quad \text{and} \quad \sigma(S^*) - \sigma(S_k) > 0 \quad \text{for all } k \in [\ell], \tag{6.9}$$

so we can write the following,

$$\begin{aligned}
&\frac{\sigma(S^*) - (1 - \beta)\sigma(S_\ell) - \beta\sigma(S_{\ell+1})}{\sigma(S^*)} \\
&\leq \frac{\sigma(S^*) - (1 - \beta)\sigma(S_\ell) - \beta\sigma(S_{\ell+1})}{\sigma(S^*|\emptyset)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma(S^*) - \sigma(S_\ell) - \beta\sigma(i_{\ell+1}|S_\ell)}{\sigma(S^*) - \sigma(S_\ell)} \prod_{k\in[\ell]} \frac{\sigma(S^*) - \sigma(S_k)}{\sigma(S^*) - \sigma(S_{k-1})} \\
&= \left(1 - \frac{\beta\sigma(i_{\ell+1}|S_\ell)}{\sigma(S^*) - \sigma(S_\ell)}\right) \prod_{k\in[\ell]} \left(1 - \frac{\sigma(i_k|S_{k-1})}{\sigma(S^*) - \sigma(S_{k-1})}\right) \\
&\leq \left(1 - \frac{\beta c_{i_{\ell+1}}}{\mathbf{e}_{S^*}^\intercal \mathbf{c}}\right) \prod_{k\in[\ell]} \left(1 - \frac{c_{i_k}}{\mathbf{e}_{S^*}^\intercal \mathbf{c}}\right) && \text{(6.7) and (6.9)} \\
&\leq \exp\left(-\frac{\beta c_{i_{\ell+1}} + \sum_{k\in[\ell]} c(i_k)}{\mathbf{e}_{S^*}^\intercal \mathbf{c}}\right) && 1 - x \leq e^{-x} \\
&= \exp\left(-\frac{(1-\beta)\mathbf{e}_{S_\ell}^\intercal \mathbf{c} + \beta \mathbf{e}_{S_{\ell+1}}^\intercal \mathbf{c}}{\mathbf{e}_{S^*}^\intercal \mathbf{c}}\right) \\
&= e^{-1}. && \text{(6.8)}
\end{aligned}
$$

Rearranging the inequality, we obtain $(1 - e^{-1})\sigma(S^*) \leq (1-\beta)\sigma(S_\ell) + \beta\sigma(S_{\ell+1})$, i.e.,

$$
\left(1 - e^{-1}\right) \frac{\sigma(S^*)}{\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}} \leq \frac{(1-\beta)\sigma(S_\ell) + \beta\sigma(S_{\ell+1})}{(1-\beta)\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c} + \beta \mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}}.
$$

The output S of Algorithm 10 maximizes the ratio of $\sigma(S_k)/\mathbf{e}_{S_k\cup\{0\}}^\intercal \mathbf{c}$ over k. Thus,

$$
\max_{k\leq\ell+1} \frac{\sigma(S_k)}{\mathbf{e}_{S_k\cup\{0\}}^\intercal \mathbf{c}} \leq \frac{\sigma(S)}{\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}}.
$$

We end the proof remarking that

$$
\max_{k\in\{\ell,\ell+1\}} \frac{\sigma(S_k)}{\mathbf{e}_{S_k\cup\{0\}}^\intercal \mathbf{c}} \geq \frac{(1-\beta)\sigma(S_\ell) + \beta\sigma(S_{\ell+1})}{(1-\beta)\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c} + \beta \mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}}.
$$

□

Notice, a similar result as Proposition 20 is stated in Theorem 3.2 of Bai et al. (2016). However, their proof doesn't hold in our case, since their inequality (16) would be true only for a *normalized* cost (i.e. $c_0 = 0$). Actually, $c_0 = 0$ implies that S^* is a singleton, from subadditivity of σ.

For more efficiency, we use a greedy algorithm with lazy evaluations (Minoux, 1978; Leskovec, Krause, et al., 2007), leveraging on the *submodularity* of σ. More precisely, in Algorithm 10, instead of taking the arg max in the step

$$
S_k \leftarrow S_{k-1} \cup \left\{ \operatorname*{arg\,max}_{i\in V\setminus S_{k-1}} \frac{\sigma(\{i\} \cup S_{k-1}) - \sigma(S_{k-1})}{c_i} \right\},
$$

we maintain an upper bound ρ (initially ∞) on hhe marginal gain, sorted in decreasing order. In each iteration k, we evaluates the element on top of the list, say i, and updates its upper bound with the marginal gain at S_{k-1}. If after the update the upper bound is greater than the others, submodularity guarantees that i is the element with the largest marginal gain.

Algorithm 10 (and the approximation factor) can't be used directly in the OIM context, since computing the exact spread σ is #P hard (Chen, Wang, and Wang, 2010). However, with Monte Carlo (MC) simulations, it can efficiently reach an

Algorithm 10 GREEDY for ratio, Lazy implementation

Input: σ that is an increasing submodular function,
$\mathbf{c} \in [0,1]^{V \cup \{0\}}$.
$S_0 \leftarrow \emptyset$.
$\rho \leftarrow [(\infty, i)]_{i \in V}$.
for $k \in [|V|]$ **do**
  `checked` $\leftarrow S_{k-1}$.
  $(*)$ Remove the first element $\rho[0] = (\sim, i)$ from ρ.
  if $i \notin$ `checked` **then**
    Insert $((\sigma(\{i\} \cup S_{k-1}) - \sigma(S_{k-1}))/c_i, i)$ in ρ, such that $\rho[:][0]$ is sorted in decreasing order.
    Add i to `checked` and go back to $(*)$.
  else
    $S_k \leftarrow S_{k-1} \cup \{i\}$.
  end if
end for
$k' \leftarrow \arg\max_{k \in \{0,\dots,|V|\}} \sigma(S_k)/\mathbf{e}_{S_k \cup \{0\}}^{\mathsf{T}} \mathbf{c}$.
Output: $S_{k'}$.

arbitrarily close ratio of $\alpha = 1 - 1/e - \varepsilon$, with a high probability $1 - 1/\left(t \log^2(t)\right)$ (Kempe, Kleinberg, and Tardos, 2015).

6.2.4 Regret bound for Algorithm 9

We provide a gap dependent upper bound on the regret of BOIM-CUCB in Theorem 27. For this, we define, for $i \in V$, the gap

$$\Delta_{i,\min} \triangleq \min_{S \subset V,\ p_i(S;\mathbf{w}^*)>0,\ \Delta(S)>0} \Delta(S).$$

We also define, with d_k being the out-degree of node k,

$$p_{i,\max} \triangleq \max_{S \subset V,\ p_i(S;\mathbf{w}^*)>0} \sum_{k \in V} d_k p_k(S;\mathbf{w}^*).$$

We also state the following smoothness property of the spread.

Fact 7 (Smoothness property of the spread)**.** *for all $S \subset V$, and all $\mathbf{w}, \mathbf{w}' \in [0,1]^E$,*

$$\forall k \in V,\ \left|p_k(S;\mathbf{w}) - p_k(S;\mathbf{w}')\right| \leq \sum_{ij \in E} p_i(S;\mathbf{w}) \left|w_{ij} - w'_{ij}\right|.$$

In particular,

$$\left|\sigma(S;\mathbf{w}) - \sigma(S;\mathbf{w}')\right| \leq |V| \sum_{ij \in E} p_i(S;\mathbf{w}) \left|w_{ij} - w'_{ij}\right|.$$

We refer the reader to Proposition 22 for a proof of the above fact, where we provide a more general statement.

Theorem 27. *If π is the policy described in Algorithm 9, then*

$$R_{B,\varepsilon}(\pi) = \mathcal{O}\left(\log B \left(\sum_{i \in V} |V| \frac{\lambda^* + d_i p_{i,\max} |V|}{\Delta_{i,\min}}\right)\right).$$

In addition, if true costs are known, then

$$R_{B,\varepsilon}(\pi) = \mathcal{O}\left(\log B\left(\sum_{i\in V}\frac{d_i p_{i,\max}|V|^2}{\Delta_{i,\min}}\right)\right).$$

Proof. Let $\alpha = 1 - 1/e - \varepsilon$, and $t \geq 1$. From Proposition 19, we have to upper bound

$$\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1}\Delta(S_t)\right].$$

Fix $t \geq 1$. We consider the following events:

$$\mathfrak{W}_t \triangleq \left\{\forall ij \in E,\ 0 \leq w^*_{ij} - w_{ij,t} \leq 2\sqrt{\frac{1.5\log(t)}{N_{\oplus ij,t-1}}}\right\},$$

$$\mathfrak{C}_t \triangleq \left\{\forall i \in V \cup \{0\},\ 0 \leq c^*_i - c_{i,t} \leq 2\sqrt{\frac{1.5\log(t)}{N_{\ominus i,t-1}}}\right\}.$$

We also consider the event $\mathfrak{A}_t$ under which the $\alpha-$approximation in Algorithm 9 holds. We already saw that

$$\mathbb{P}[\neg\mathfrak{A}_t] \leq \frac{1}{t\log^2(t)}.$$

From Hoeffding's inequality, $\mathfrak{C}_t$ holds with probability $1 - 2(|V|+1)/t^2$, and $\mathfrak{W}_t$ holds with probability $1 - 2|E|/t^2$. Thus, under the event that either $\mathfrak{C}_t$, $\mathfrak{W}_t$ or $\mathfrak{A}_t$ doesn't hold, then the regret is bounded by a constant (since we have a convergent series).

We thus now assume that $\mathfrak{C}_t$, $\mathfrak{W}_t$ and $\mathfrak{A}_t$ hold. In particular, using $\mathfrak{C}_t$ and $\mathfrak{W}_t$, we can write

$$\lambda^* = \frac{\sigma(S^*;\mathbf{w}^*)}{\mathbf{e}^\intercal_{S^*\cup\{0\}}\mathbf{c}^*} \leq \frac{\sigma(S^*;\mathbf{w}_t)}{\mathbf{e}^\intercal_{S^*\cup\{0\}}\mathbf{c}_t}.$$

We can use this relation to write

$$\begin{aligned}
\Delta(S_t) &= \lambda^*\alpha(\mathbf{e}^\intercal_{S_t}\mathbf{c}^* + c^*_0) - \sigma(S_t;\mathbf{w}^*)\\
&= \lambda^*\alpha\left((\mathbf{e}^\intercal_{S_t}\mathbf{c}^* + c^*_0) - \frac{\sigma(S_t;\mathbf{w}_t)}{\alpha\lambda^*}\right) + (\sigma(S_t;\mathbf{w}_t) - \sigma(S_t;\mathbf{w}^*))\\
&\leq \lambda^*\alpha\left((\mathbf{e}^\intercal_{S_t}\mathbf{c}^* + c^*_0) - \frac{\sigma(S_t;\mathbf{w}_t)}{\alpha\sigma(S^*;\mathbf{w}_t)}(\mathbf{e}^\intercal_{S^*}\mathbf{c}_t + c_{0,t})\right) + (\sigma(S_t;\mathbf{w}_t) - \sigma(S_t;\mathbf{w}^*))\\
&\leq \lambda^*\alpha((\mathbf{e}^\intercal_{S_t}\mathbf{c}^* + c^*_0) - (\mathbf{e}^\intercal_{S_t}\mathbf{c}_t + c_{0,t})) + (\sigma(S_t;\mathbf{w}_t) - \sigma(S_t;\mathbf{w}^*)),
\end{aligned}$$

where the last inequality is from $\mathfrak{A}_t$. Notice here that in the case the costs are known, the first term in this bound disappears, and we can then safely take $\lambda^* = 0$, explaining why in the final bound the term in front of λ^* disappears. We now use Fact 7, and then $\mathfrak{C}_t$, $\mathfrak{W}_t$ to further get the bound

$$\Delta(S_t) \leq \underbrace{\lambda^*\alpha\sum_{i\in S_t}\left(1 \wedge 2\sqrt{\frac{1.5\log(t)}{N_{\ominus i,t-1}}}\right)}_{(6.10)} + \underbrace{|V|\sum_{ij\in E}p_i(S_t;\mathbf{w}^*)\left(1 \wedge 2\sqrt{\frac{1.5\log(t)}{N_{\oplus ij,t-1}}}\right)}_{(6.11)}.$$

Then, necessarily either $\Delta(S_t) \leq 2\cdot(6.10)$ or $\Delta(S_t) \leq 2\cdot(6.11)$ is true. The first event can be handle with Theorem 12 (with no probabilistically triggered arms),

using the upper bound on the expectation of the random horizon given in the proof of Theorem 19. This allows us to get a term of order

$$\lambda^* \log(B/c_0^*) \sum_{i \in V} \frac{|V|}{\Delta_{i,\min}},$$

in the regret upper bound.

Theorem 12 with probabilistically triggered arms takes care of the second event to get a bound of order

$$\log(B/c_0^*) \sum_{i \in V} d_i \frac{|V|^2 \max_{S \subset V,\ p_i(S;\mathbf{w}^*)>0} \sum_{k \in V} d_k p_k(S;\mathbf{w}^*)}{\Delta_{i,\min}}.$$

Problem-independent bound The problem-independent bound of

$$\mathcal{O}\left(|V| \sqrt{B \log B \sum_{i \in V} d_i p_{i,\max}}\right)$$

is an immediate consequence of our problem-dependent bound, decomposing, classically, the regret in two terms by filtering by whether or not $\Delta(S_t) \leq \delta$, and then taking the worst regime for δ. □

Notice that the analysis can be easily used for the non budgeted setting. In this case, it reduces to the state-of-the-art analysis of Wang and Chen (2017), except that we slightly simplify and improve the analysis to replace the factor

$$\max_{S \subset V} \sum_{k \in V} d_k \mathbb{I}\{p_k(S;\mathbf{w}^*) > 0\}$$

by a potentially much lower quantity $p_{i,\max}$. In the case this last quantity is still large, we can further improve it by considering slight modifications to the original Algorithm 9. This is the purpose of the next section.

6.3 ℓ_2-based approach

We observe that the factor $p_{i,\max}$ in Theorem 27 can be as large as $|E|$ in the worst case. In other word, if $\Delta = \min_i \Delta_{i,\min}$, the rate can be as large as

$$\mathcal{O}\left(\log B \left(\frac{\lambda^* |V|^2 + |E|^2 |V|^2}{\Delta}\right)\right).$$

We argue here that we can replace $|E|^2|V|^2$ by $|E||V|^3 \log^2(|V|))$. Indeed, leveraging on the mutual independence of random variables W_{ij}, we can hope to get a tighter confidence region for $\mathbf{w}^*$, and thus a provably tighter regret upper bound (Magureanu, Combes, and Proutiere, 2014; Combes et al., 2015; Degenne and Perchet, 2016b). We consider the following ℓ_2 confidence region:

Fact 8 (Confidence ellipsoid for weights, Degenne and Perchet (2016b)). *For all $t \geq 2$, with probability at least $1 - 1/\left(t \log^2(t)\right)$,*

$$\sum_{ij \in E} N_{\oplus i,t-1} \left(w_{ij}^* - \overline{w}_{ij,t-1}\right)^2 \leq \delta(t),$$

where $\delta(t) \triangleq 2\log(t) + 2(|E|+2)\log\log(t) + 1$.

For OIM (both budgeted and non budgeted), there is a large potential gain in the analysis using the confidence region given by Fact 8 compared to simply using an Hoeffding based one, like in BOIM-CUCB. More precisely, for classical combinatorial semi bandits, Degenne and Perchet (2016b) reduced the gap dependent regret upper bound by a factor $\ell/\log^2(\ell)$, where in our case ℓ can be as large as $|E|$. However, there is also a drawback in practice with such confidence region: computing the optimistic spread might be inefficient, even if an oracle for evaluating the spread is available. Indeed, for a fixed $S \subset V$, the problem of maximizing $\mathbf{w} \mapsto \sigma(S;\mathbf{w})$ over $\mathbf{w}$ belonging to some ellipsoid might be hard, since the objective is not necessarily concave. We can overcome this issue using the Fact 7. For $S \subset V$ and $\mathbf{w} \in \mathbb{R}^E$, we define the confidence "bonus" as follows:

$$\text{bonus}(S;\mathbf{w}) \triangleq |V| \sqrt{\delta(t) \sum_{i \in V, N_{\oplus i,t-1}>0} d_i \frac{p_i(S;\mathbf{w})^2}{N_{\oplus i,t-1}}}.$$

Notice, we don't sum on vertices with a zero counter. We compensate this by using the convention $\overline{w}_{ij,t-1} = 1$ when $N_{\oplus i,t-1} = 0$. We can successively use Fact 7, Cauchy-Schwartz inequality, and Fact 8 to get, with probability at least $1 - 1/\left(t\log^2(t)\right)$,

$$\sigma(S;\mathbf{w}^*) \leq \sigma(S;\overline{\mathbf{w}}_{t-1}) + \text{bonus}(S;\overline{\mathbf{w}}_{t-1}). \tag{6.12}$$

In the same way, with probability at least $1 - 1/\left(t\log^2(t)\right)$, we also have (6.13).

$$\sigma(S;\mathbf{w}^*) \leq \sigma(S;\overline{\mathbf{w}}_{t-1}) + \text{bonus}(S;\mathbf{w}^*). \tag{6.13}$$

Contrary to (6.12), this "optimistic spread" can't be used directly by the agent since $\mathbf{w}^*$ is not known.

Although the optimistic spread defined in (6.12) is now much easier to compute, there is still a major drawback that remains: As a function of $S \subset V$, $\text{bonus}(S;\overline{\mathbf{w}}_{t-1})$ is not necessarily submodular, so the optimistic spread is itself no longer submodular. This is an issue because submodularity is a crucial property for reaching the approximation ratio $1 - 1/e - \varepsilon$. We propose here several submodular upper bound to bonus, defined for $S \subset V$ and $\mathbf{w} \in \mathbb{R}^E$:

- bonus_1 is actually modular, and simply uses the subadditivity (w.r.t. S) of bonus:
$$\text{bonus}_1(S;\mathbf{w}) \triangleq |V| \sum_{j \in S} \sqrt{\delta(t) \sum_{i, N_{\oplus i,t-1}>0} d_i \frac{p_i(\{j\};\mathbf{w})^2}{N_{\oplus i,t-1}}}.$$
- bonus_2 uses the subadditivity of the square root:
$$\text{bonus}_2(S;\mathbf{w}) \triangleq |V| \sum_{i, N_{\oplus i,t-1}>0} p_i(S;\mathbf{w}) \sqrt{\frac{\delta(t) d_i}{N_{\oplus i,t-1}}}.$$
- bonus_3 uses $p_i(S;\mathbf{w})^2 \leq p_i(S;\mathbf{w})$, and is submodular as the composition between a non decreasing concave function (the square root) and a monotone

submodular function:

$$\text{bonus}_3(S;\mathbf{w}) \triangleq |V|\sqrt{\delta(t)\sum_{i,N_{\oplus\, i,t-1}>0} d_i \frac{p_i(S;\mathbf{w})}{N_{\oplus i,t-1}}}.$$

- bonus_4 uses Jensen's inequality, and is submodular as the expectation of the square root of a submodular function.

$$\text{bonus}_4(S;\mathbf{w}) \triangleq \mathbb{E}\left[|V|\sqrt{\sum_{i\in V, N_{\oplus\, i,t-1}>0, S\overset{\mathbf{W}}{\rightsquigarrow} i} \frac{\delta(t) d_i}{N_{\oplus i,t-1}}}\right],$$

where $\mathbf{W} \sim \otimes_{ij\in E}\text{Bernoulli}(w_{ij})$.

We can write the following approximation guarantees for the two first bonus:

$$\text{bonus}(S;\mathbf{w}) \leq \text{bonus}_1(S;\mathbf{w}) \leq |S|\text{bonus}(S;\mathbf{w}), \tag{6.14}$$

$$\text{bonus}(S;\mathbf{w}) \leq \text{bonus}_2(S;\mathbf{w}) \leq \sqrt{|V|}\text{bonus}(S;\mathbf{w}).$$

Notice, another approach to get a submodular bonus is to approximate $p_i(S;\overline{\mathbf{w}}_{t-1})$ by the square root of a modular function (Goemans et al., 2009). However, not only this bonus would be much more computationally costly to build than ours, but also, we would get only a $\sqrt{|V|}\log|V|$ approximation factor, which is worst than the one with our bonus_2. Since increasing the bonus by a factor $\alpha \geq 1$ increases the gap dependent regret upper bound by a factor α^2, we only loose a factor $|V|$ for bonus_2, compared to the use of bonus, which is still better than the CUCB approach. bonus_1 can also be interesting to use when we have some upper bound guarantee on the cardinality of seed sets used. An approximation factor for bonus_3 or bonus_4 doesn't seem interesting, because it would involve the inverse of triggering probabilities. We can, however, further upper bound $\text{bonus}_3(S;\mathbf{w}^*)$ as follows:

$$\begin{aligned}
\text{bonus}_3(S;\mathbf{w}^*) &= |V|\sqrt{\delta(t)\sum_{i,N_{\oplus\, i,t-1}>0} d_i \frac{p_i(S;\mathbf{w}^*)}{N_{\oplus i,t-1}}} \\
&\leq |V|\sqrt{\sum_{j\in S}\sum_{\substack{i\in V,\\ N_{\oplus\, i,t-1}>0}} d_i \frac{\delta(t) p_i(\{j\};\mathbf{w}^*)}{N_{\oplus i,t-1}}} \\
&\leq |V|\sqrt{\delta(t)\sum_{j\in S}|E|\left(\frac{8}{N_{\ominus j,t-1}} \wedge 1\right)},
\end{aligned}$$

where the last inequality only holds under some high probability event, given by the following Proposition 21 (which is proven in subsection 6.5.2), involving counters on the costs and counters on the weights.

Proposition 21. *Consider the event defined by* $\mathfrak{P}_t \triangleq \{\forall i \in V, N_{\oplus\, i,t-1} \geq \delta(t)\}$. *Then, for all* $i, j \in V$,

$$\mathbb{P}\left[\mathfrak{P}_t \text{ and } \frac{\delta(t) p_i(\{j\};\mathbf{w}^*)}{N_{\oplus i,t-1}} > \frac{8\delta(t)}{N_{\ominus j,t-1}}\right] \leq 1/t^2.$$

We thus define for $S \subset V$,

$$\text{bonus}_5(S) \triangleq |V| \sqrt{\delta(t) \sum_{j \in S} |E| \left(\frac{8}{N_{\ominus j,t-1}} \wedge 1 \right)}.$$

This bonus is much more convenient since it does not depend anymore on $\mathbf{w}^*$, and can thus be computed by the agent. Indeed, although the first four bonuses are likely to be tighter than this last submodular bonus_5, their dependence in $\mathbf{w}$ forces us to use them for $\mathbf{w} = \overline{\mathbf{w}}_{t-1}$. Even if this doesn't pose any problem in practice, this is more difficult to handle in theory since it would involve optimistic estimates on $p_i(\,\cdot\,; \overline{\mathbf{w}}_{t-1})$ itself (see the next section for further details). Actually, we will see that the analysis with bonus_5 is slightly better than the one we would get with bonus_2, since we loose a factor $|V| \log^2(|V|) / \log^2(|E|)$ compared to the use of bonus. In addition, it allows a much more interesting constant term in the regret upper bound thanks to the suppression of the dependence in $\mathbf{w}^*$.

Although we can improve the analysis based on Fact 8 and 7, the inequality in this last fact may be be less rough in practice (we confirm this in section 6.4). In that case, we suffer from this roughness, since we actually use Fact 7 to design our bonus. In contrast, BOIM-CUCB only uses it in the analysis, and can therefore adapt to a better smoothness inequality. Thus, we consider BOIM-CUCB$_5$, where we first compute S_t using the BOIM-CUCB approach, and accept it only if

$$\sigma(S_t; \mathbf{w}_t) \leq \sigma(S_t; \overline{\mathbf{w}}_{t-1}) + \text{bonus}_5(S_t), \tag{6.15}$$

otherwise, we chose S_t maximizing $\sigma(S; \overline{\mathbf{w}}_{t-1}) + \text{bonus}_5(S)$. For technical reason (due to Proposition 21), we replace S_t by $S_t \cup \{j\}$ if it exists a $j \in V$, such that

$$N_{\oplus j,t-1} < \delta(t). \tag{6.16}$$

We thus both enjoy the theoretical advantages of bonus_5 and the practical advantages of BOIM-CUCB. We give the following regret bounds for this approach.

Theorem 28. *If π is the policy following BOIM-CUCB$_5$, then*

$$R_{B,\varepsilon}(\pi) = \mathcal{O}\left(\log B \left(\sum_{i \in V} |V| \frac{\lambda^* + |V||E| \log^2 |V|}{\Delta_{i,\min}} + \lambda^* |V|^2 \right) \right).$$

Proof. Let $\alpha = 1 - 1/e - \varepsilon$, and $t \geq 1$. From Proposition 19, we have to upper bound

$$\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \Delta(S_t) \right].$$

In the proof, in addition to $\mathfrak{P}_t \triangleq \{\forall i \in V, N_{\oplus i,t-1} \geq \delta(t)\}$, we consider the following events:

$$\mathfrak{W}_t \triangleq \left\{ \sum_{ij \in E} N_{\oplus i,t-1} \left(w^*_{ij} - \overline{w}_{ij,t-1} \right)^2 \leq 2\delta(t) \right\},$$

$$\mathfrak{C}_t \triangleq \left\{ \forall i \in V \cup \{0\},\ 0 \leq c^*_i - c_{i,t} \leq 1 \wedge 2\sqrt{\frac{1.5 \log(t)}{N_{\ominus i,t-1}}} \right\}.$$

$$\mathfrak{B}_t \triangleq \left\{ \forall i, j \in V,\ \frac{\delta(t) p_i(\{j\}; \mathbf{w}^*)}{N_{\oplus i,t-1}} \leq \frac{8\delta(t)}{N_{\ominus j,t-1}} \right\}.$$

For each node $i \in V$, if $N_{\oplus i,t-1} \geq \delta(\tau_B - 1)$, then i will not be intentionally added to the seed set in BOIM-CUCB$_5$. Then, each node is intentionally added for at most $\delta(\tau_B - 1) + 1$ times. Thus, we can write

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \Delta(S_t)\mathbb{I}\{\neg\mathfrak{P}_t\}\right] &\leq \mathbb{E}[(\delta(\tau_B - 1) + 1)|V|\lambda^*(|V| + 1)] \\ &\leq \Big(\delta\Big((2B/c_0^* + 1)^2\Big) + 1\Big)|V|\lambda^*(|V| + 1).\end{aligned}$$

We can therefore assume that $\mathfrak{P}_t$ holds. In this case, we have by Proposition 21 that $\mathfrak{B}_t$ doesn't hold with probability bounded by $|V|^2/t^2$. On the other hand, from Fact 8, $\mathfrak{W}_t$ doesn't hold with probability bounded by $1/\Big(t\log^2(t)\Big)$, and from Hoeffding inequality, $\mathfrak{C}_t$ doesn't hold with probability bounded by $2(|V| + 1)/t^2$. We can consider the event $\mathfrak{A}_t$ under which the $\alpha-$approximation in BOIM-CUCB$_5$ holds. We already saw that

$$\mathbb{P}[\neg\mathfrak{A}_t] \leq \frac{1}{t\log^2(t)}.$$

The regret in the case one of the events $\neg\mathfrak{A}_t, \neg\mathfrak{B}_t, \neg\mathfrak{W}_t, \neg\mathfrak{C}_t$ holds is thus bounded by a constant depending on $|V|$ and λ^*. It thus remains to upper bound

$$\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \Delta(S_t)\mathbb{I}\{\mathfrak{A}_t, \mathfrak{B}_t, \mathfrak{W}_t, \mathfrak{C}_t\}\right].$$

For this, notice that from $\mathfrak{A}_t$, S_t which is the seed set chosen by our policy at round t, is an α-approximate maximizer of $A \mapsto f(A)/(\mathbf{e}_A^\intercal \mathbf{c}_t + c_{0,t})$, where f is one of the optimistic spreads considered in BOIM-CUCB$_5$. We thus have

$$\frac{f(S_t)}{\mathbf{e}_{S_t}^\intercal \mathbf{c}_t + c_{0,t}} \geq \alpha \frac{f(S^*)}{\mathbf{e}_{S^*}^\intercal \mathbf{c}_t + c_{0,t}},$$

where $S^* \in \arg\max_{S\subset V} \frac{\sigma(S;\mathbf{w}^*)}{\mathbf{e}_S^\intercal \mathbf{c}^* + c_0^*}$. Since under $\mathfrak{W}_t$, $f(S^*) \geq \sigma(S^*;\mathbf{w}^*)$, we can derive the following upper bound on the gap:

$$\begin{aligned}\Delta(S_t) &= \lambda^*\alpha(\mathbf{e}_{S_t}^\intercal \mathbf{c}^* + c_0^*) - \sigma(S_t;\mathbf{w}^*) \\ &= \lambda^*\alpha\left((\mathbf{e}_{S_t}^\intercal \mathbf{c}^* + c_0^*) - \frac{f(S_t)}{\alpha\lambda^*}\right) + (f(S_t) - \sigma(S_t;\mathbf{w}^*)) \\ &\leq \lambda^*\alpha\left((\mathbf{e}_{S_t}^\intercal \mathbf{c}^* + c_0^*) - \frac{f(S_t)}{\alpha f(S^*)}(\mathbf{e}_{S^*}^\intercal \mathbf{c}_t + c_{0,t})\right) + (f(S_t) - \sigma(S_t;\mathbf{w}^*)) \quad \mathfrak{W}_t, \mathfrak{C}_t \\ &\leq \lambda^*\alpha((\mathbf{e}_{S_t}^\intercal \mathbf{c}^* + c_0^*) - (\mathbf{e}_{S_t}^\intercal \mathbf{c}_t + c_{0,t})) + (f(S_t) - \sigma(S_t;\mathbf{w}^*)). \quad \mathfrak{A}_t\end{aligned}$$

From this point, we can use the condition satisfied by f in BOIM-CUCB$_5$:

$$f(S_t) \leq \sigma(S_t;\overline{\mathbf{w}}_{t-1}) + \text{bonus}_5(S_t).$$

Using Fact 7 with Fact 8, we can further have with Cauchy-Schwartz inequality

$$\sigma(S_t;\overline{\mathbf{w}}_{t-1}) - \sigma(S_t;\mathbf{w}^*) \leq \text{bonus}_5(S_t).$$

This allows us to get, using $\mathfrak{C}_t$,

$$\Delta(S_t) \leq \lambda^* \alpha \sum_{i \in S_t \cup \{0\}} 1 \wedge 2\sqrt{\frac{1.5\log(t)}{N_{\ominus i,t-1}}} + 2\text{bonus}_5(S_t). \tag{6.17}$$

The first part in (6.17) can be handle using Theorem 12 to get a term of order

$$\lambda^* \log(B/c_0^*) \sum_{i \in V} \frac{|V|}{\Delta_{i,\min}},$$

in the regret upper bound. We can use Theorem 13 to deal with the second part, to get a term of order

$$\delta(B/c_0^*)|V|^2|E| \sum_{i \in V} \frac{\log^2(|V|)}{\Delta_{i,\min}},$$

in the regret upper bound. We thus get the desired result. □

Such analysis also holds in the non-budgeted setting, and maximizing the spread only instead of the ratio, we can build a policy π satisfying the following (with the standard definition of the non-budgeted gaps):

$$R_{T,\varepsilon}(\pi) = \mathcal{O}\left(\log T \sum_{i \in V} \frac{|V|^2|E|\log^2|V|}{\Delta_{i,\min}}\right).$$

The regret rate is thus better than the one from Wang and Chen (2017), gaining a factor $|E|/\Big(|V|\log^2(|V|)\Big)$.

6.3.1 Improvements using bonus_1 and bonus_4

In this subsection, we show that the use of bonus_1 and bonus_4 leads to a better regret leading term, at the cost of a large second order term. In the following, we propose BOIM-CUCB$_1$ (resp. BOIM-CUCB$_4$), that are the same approach as BOIM-CUCB$_5$ with $\text{bonus}_1(\,\cdot\,;\overline{\mathbf{w}}_{t-1})$ (resp. $\text{bonus}_4(\,\cdot\,;\overline{\mathbf{w}}_{t-1})$) instead of bonus_5, and where condition (6.16) is replaced by

$$\exists j \in V, N_{\oplus j,t-1} \leq |E|\delta(t).$$

bonus_1 for low cardinality seed sets

In many real world scenarios, maximal cardinality of seed set is small compared to $|V|$. Indeed, in the non-budgeted setting, it is limited by m, and it is usually assumed that m is much smaller than $|V|$. In the budgeted setting, we will see in subsection 6.3.2 how to limit the cost of the chosen seeds, and this is likely to also induce a limit on the cardinality of seeds. Using bonus_1 is more appropriate in this situation, according to the approximation factor (6.14). We state in Theorem 29 the regret bound for BOIM-CUCB$_1$.

Theorem 29. *If π is the policy* BOIM-CUCB$_1$*, and if all seeds selected have a cardinality bounded by m, then we have*

$$R_{B,\varepsilon}(\pi) = \mathcal{O}\left(\log B\left(\sum_{i \in V} m\frac{\lambda^* + m|V|^2 d_i \log^2(|E|)}{\Delta_{i,\min}} + \lambda^*|V|^2|E|\right)\right).$$

Proof. Let $\alpha = 1 - 1/e - \varepsilon$, and $t \geq 1$. The beginning of the proof is the same as in Theorem 28, except we no longer consider the event $\mathfrak{B}_t$, and we consider a new event:

$$\mathfrak{R}_t \triangleq \{\forall i \in V, N_{\oplus i,t-1} \geq |E|\delta(t)\}.$$

As for Theorem 28:

- The regret in the case $\mathfrak{R}_t$ doesn't hold can be bounded by a term of order

$$\lambda^* |V|^2 |E| \log(B).$$

- When all the events hold, the same analysis gives

$$\Delta(S_t) \leq 2\lambda^* \alpha \sum_{i \in S_t} 1 \wedge 2\sqrt{\frac{1.5\log(t)}{N_{\ominus i,t-1}}} + 4\text{bonus}_1(S_t; \overline{\mathbf{w}}_{t-1}),$$

  and the first term can be handled in the same way.

The second term can be analyzed in the following way: After bounding it by

$$4m\text{bonus}(S_t; \overline{\mathbf{w}}_{t-1}),$$

see that using Fact 7 on the quantity $p_i(S; \overline{\mathbf{w}}_{t-1})$ present in this bonus, we get

$$p_i(S_t; \overline{\mathbf{w}}_{t-1}) \leq p_i(S_t; \mathbf{w}^*) + \frac{1}{|V|}\text{bonus}(S_t; \mathbf{w}^*).$$

By subadditivity, and from $\mathfrak{R}_t$, we have

$$\begin{aligned}
4m\text{bonus}(S_t; \overline{\mathbf{w}}_{t-1}) &\leq 4m\text{bonus}(S_t; \mathbf{w}^*) + 4m|V|\sqrt{\delta(t)\sum_{i \in V} d_i \frac{\text{bonus}(S_t; \mathbf{w}^*)^2}{|V|^2 N_{\oplus i,t-1}}} \\
&\leq 4m\text{bonus}(S_t; \mathbf{w}^*) + 4m|V|\sqrt{\sum_{i \in V} d_i \frac{\text{bonus}(S_t; \mathbf{w}^*)^2}{|V|^2 |E|}} \\
&= 8m\text{bonus}(S_t; \mathbf{w}^*).
\end{aligned}$$

We can now use Theorem 13 to get a bound of order

$$\delta(B/c_0^*) m^2 |V|^2 \sum_{i \in V} d_i \frac{\log^2(|E|)}{\Delta_{i,\min}}.$$

□

As previously, we can state the following non-budgeted version, with seed set cardinality constrained by m:

$$R_{T,\varepsilon}(\pi) = \mathcal{O}\left(\log T \left(\sum_{i \in V} \frac{m^2 |V|^2 d_i \log^2(|E|)}{\Delta_{i,\min}} + |E||V|^2\right)\right).$$

Notice, for both settings, there is an improvement in the main term (the gap dependent one), in the case $m \leq \sqrt{|V|}$. However, there is also a higher gap independent term that appears.

bonus$_4$: the same performance as bonus?

We show here that the regret with bonus$_4$ is of the same order as what we would have had with bonus (which is not submodular). However, bonus$_4$ does not have the calculation guarantees of the other bonuses. We state in Theorem 30 the regret bound for the policy BOIM-CUCB$_4$. Notice that we obtain a bound whose leading term improves by a factor $|E|/\log^2|E|$ that of BOIM-CUCB.

Theorem 30. *If π is the policy BOIM-CUCB$_4$, then we have*

$$R_{B,\varepsilon}(\pi) = \mathcal{O}\left(\log B\left(\sum_{i\in V}\frac{|V|\lambda^* + |V|^2 d_i \log^2(|E|)}{\Delta_{i,\min}} + \lambda^*|V|^2|E|\right)\right).$$

Before proving this result, we provide some preliminaries. If we let $p_{S\rightsquigarrow S'}(\mathbf{w}) \triangleq \mathbb{P}\left[\left\{i\in V,\ S\overset{\mathbf{W}}{\rightsquigarrow} i\right\} = S'\right]$, then another expression is

$$\begin{aligned}\text{bonus}_4(S;\mathbf{w}) &= \sum_{S'\supset S} p_{S\rightsquigarrow S'}(\mathbf{w})\,\underbrace{|V|\sqrt{\delta(t)\sum_{i\in S'}\frac{d_i}{N_{\oplus i,t-1}}}}_{g(S')}\\ &= \sum_{k\geq 0}(g(S'_{k+1}) - g(S'_k)) \sum_{S'\notin\{S'_0,\ldots,S'_k\}} p_{S\rightsquigarrow S'}(\mathbf{w}).\end{aligned}$$

where $g(S) = g(S'_1) \leq g(S'_2) \leq \ldots$ and $S'_0 = \emptyset$.

Since this bonus shall be used with $\overline{\mathbf{w}}_{t-1}$, we need a smoothness inequality to link $p_{S\rightsquigarrow S'}(\overline{\mathbf{w}}_{t-1})$ to $p_{S\rightsquigarrow S'}(\mathbf{w}^*)$. We prove here the following such inequality.

Proposition 22. *For all $S\subset V$, all $\mathbf{w},\mathbf{w}'\in[0,1]^E$ and all collection of subsets of vertices $\mathcal{S}$, we have*

$$\left|\sum_{S'\in\mathcal{S}}(p_{S\rightsquigarrow S'}(\mathbf{w}) - p_{S\rightsquigarrow S'}(\mathbf{w}'))\right| \leq \sum_{ij\in E} p_i(S;\mathbf{w})\left|w'_{ij} - w_{ij}\right|.$$

Proof. We assume w.l.o.g. that $\mathbf{w}'\geq\mathbf{w}$. We consider the random graph $G_{\mathbf{W}} = (V,\{ij\in E, W_{ij}=1\})$, where $\mathbf{W}\sim\otimes_{ij\in E}\text{Bernoulli}(w_{ij})$. We build $G_{\mathbf{W}'}$ from $G_{\mathbf{W}}$ by adding edges ij independently with probability $\frac{w'_{ij}-w_{ij}}{1-w_{ij}}$ for each ij that is not an edge in $G_{\mathbf{W}}$. Now, see that

$$\sum_{S'\in\mathcal{S}} p_{S\rightsquigarrow S'}(\mathbf{w}) = \mathbb{P}\left[S\overset{\mathbf{W}}{\rightsquigarrow} S'\right] - \sum_{S'\notin\mathcal{S}} p_{S\rightsquigarrow S'}(\mathbf{w}),$$

where $S\overset{\mathbf{W}}{\rightsquigarrow} S'$ means $S\overset{\mathbf{W}}{\rightsquigarrow} i$ for all $i\in S'$. Thus,

$$\begin{aligned}0\leq p_{S\rightsquigarrow S'}(\mathbf{w}') - p_{S\rightsquigarrow S'}(\mathbf{w}) &= \mathbb{P}\left[S\overset{\mathbf{W}'}{\rightsquigarrow} S'\right] - \mathbb{P}\left[S\overset{\mathbf{W}}{\rightsquigarrow} S'\right]\\ &\quad - \left(\sum_{S'\notin\mathcal{S}} p_{S\rightsquigarrow S'}(\mathbf{w}') - \sum_{S'\notin\mathcal{S}} p_{S\rightsquigarrow S'}(\mathbf{w})\right)\\ &\leq \mathbb{P}\left[S\overset{\mathbf{W}'}{\rightsquigarrow} S'\right] - \mathbb{P}\left[S\overset{\mathbf{W}}{\rightsquigarrow} S'\right]\\ &= \mathbb{P}\left[S\overset{\mathbf{W}'}{\rightsquigarrow} S' \text{ but not } S\overset{\mathbf{W}}{\rightsquigarrow} S'\right]\end{aligned}$$

$$\leq \mathbb{P}\Big[\exists ij \in E \text{ s.t. } S \overset{\mathbf{W}}{\rightsquigarrow} i,\ W_{ij} = 0,\ \text{and } W'_{ij} = 1\Big].$$

The last inequality is by noticing that if $S \overset{\mathbf{W}'}{\rightsquigarrow} S'$ but not $S \overset{\mathbf{W}}{\rightsquigarrow} S'$, then there must be a edge ij accessible from S in $G_{\mathbf{W}'}$ such that $W_{ij} = 0$ and $W'_{ij} = 1$. Taking the first such edge ij (watching the contagion spread from S step by step), we see that ij must be accessible from S in $G_{\mathbf{W}}$ as well (since otherwise there's a previous accessible edge $k\ell$ that verifies $W_{k\ell} = 0$ and $W'_{k\ell} = 1$).

We have that $S \overset{\mathbf{W}}{\rightsquigarrow} i$ is independent from W_{ij}, W'_{ij}. Since

$$\mathbb{P}\Big[W_{ij} = 0,\ \text{and } W'_{ij} = 1\Big] = (1 - w_{ij})\frac{w'_{ij} - w_{ij}}{1 - w_{ij}} = w'_{ij} - w_{ij},$$

we have

$$\mathbb{P}\Big[\exists ij \in E \text{ s.t. } S \overset{\mathbf{W}}{\rightsquigarrow} i,\ W_{ij} = 0,\ \text{and } W'_{ij} = 1\Big] \leq \sum_{ij \in E} p_i(S; \mathbf{w})\Big(w'_{ij} - w_{ij}\Big).$$

□

Proof of Theorem 30. We apply a similar analysis as above. When all the events hold, the same analysis gives

$$\Delta(S_t) \leq 2\lambda^* \alpha \sum_{i \in S_t} 1 \wedge 2\sqrt{\frac{1.5\log(t)}{N_{\ominus i,t-1}}} + 4\text{bonus}_4(S_t; \overline{\mathbf{w}}_{t-1}),$$

and the first term can be handled in the same way. The second term can be analyzed in the following way: Using Proposition 22 with $\mathcal{S} = \{S' \subset V,\ S' \notin \{S'_0, \dots, S'_k\}\}$, we get

$$\begin{aligned}
4\text{bonus}_4(S_t; \overline{\mathbf{w}}_{t-1}) &\leq 4\text{bonus}_4(S_t; \mathbf{w}^*) + \sum_{k \geq 0}(g(S'_{k+1}) - g(S'_k))\frac{1}{|V|}\text{bonus}_4(S_t; \mathbf{w}^*) \\
&= \left(4 + \sqrt{\delta(t)\sum_{i \in V}\frac{d_i}{N_{\oplus i,t-1}}}\right)\text{bonus}_4(S_t; \mathbf{w}^*) \\
&\leq 5\text{bonus}_4(S_t; \mathbf{w}^*),
\end{aligned}$$

where the last inequality uses the event

$$\mathfrak{R}_t \triangleq \{\forall i \in V, N_{\oplus i,t-1} \geq |E|\delta(t)\}.$$

Relying on Theorem 14, we can deal with this last term and obtain a term of order

$$\delta B / c_0^* \sum_{i \in V} \frac{|V|^2 d_i \log^2(|E|)}{\Delta_{i,\min}}.$$

□

As previously, we can state the following non-budgeted version. Notice that the cardinality constraint does not appear in the bound.

$$R_{T,\varepsilon}(\pi) = \mathcal{O}\left(\log T\left(\sum_{i \in V}\frac{|V|^2 d_i \log^2(|E|)}{\Delta_{i,\min}} + |E||V|^2\right)\right).$$

In spite of the superiority in terms of regret of the use of bonus$_4$, we must point out that, in the worst case, the calculation of this bonus may require a number of sample (and thus a time complexity) polynomial in t, which does not meet the criterion of efficiency that we set ourselves at the beginning of the chapter.

6.3.2 Knapsack constraint for known costs

In their setting, Wang, Yang, et al. (2020) considered the relaxed constraint

$$\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}\mathbf{c}^*\right] \leq b, \tag{6.18}$$

instead of ratio maximization, where the expectation is over the possible randomness of S. When true costs are known to the agent, we can actually combine the two settings: a seed set S can be chosen only if it satisfies (6.18). In this section, we describe modifications this new setting implies. First of all, the regret definition is impacted, and F_B^* is now maximal for policies respecting the constraint (6.18) within each round. Naturally, the definitions of λ^* and S^* are also modified accordingly. Otherwise, apart from Algorithm 10, there is conceptually no change in the approaches that have been described in this paper. We now described the modification needed to make Algorithm 10 works in this setting. The same sequence of set S_k is considered, but instead of choosing the set that maximizes the ratio over all $k \in \{0, \dots, |V|\}$, we restrict the maximization to $k \in \{0, \dots, j\}$, where j is the first index such that $\mathbf{e}_{S_j\cup\{0\}}^\intercal \mathbf{c}^* > b$. If this maximizer is not S_j, then it satisfies the constraint and is output. Else, we output S_j with probability $(b - \mathbf{e}_{S_{j-1}\cup\{0\}}^\intercal \mathbf{c}^*)/c_j^*$ and S_{j-1} with probability $1 - (b - \mathbf{e}_{S_{j-1}\cup\{0\}}^\intercal \mathbf{c}^*)/c_j^*$. This way, the expected cost of the output is b. The following Proposition 23 gives an approximation factor of $1 - 1/e$ for the above modification of Algorithm 10.

Proposition 23. *The solution S obtained by the modified Algorithm 10 is such that:*

$$\left(1 - e^{-1}\right) \frac{\mathbb{E}[\sigma(S^*)]}{\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right]} \leq \frac{\mathbb{E}[\sigma(S)]}{\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^*\right]},$$

where the expectation is over the possible randomness of S, S^*.

Proof. There are two possibilities for S^*: either $\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right] < b$, or $\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right] = b$. In the first case, we know that S^* is not random. Indeed, if it is not the case, then $\frac{\mathbb{E}[\sigma(S^*)]}{\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right]}$ is a convex combination of some $\frac{\sigma(S)}{\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^*}$ for S in the support of the distribution of S^*. Necessarily, the maximizer (over S in the support) of the ratio is such that $\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^* > b$, since otherwise this maximizer contradicts the definition of S^*. Therefore, increasing the coefficient of this maximizer in the convex combination increases $\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right]$, which can thus be set to b. Since this also increases $\frac{\mathbb{E}[\sigma(S^*)]}{\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right]}$, we get a contradiction since we improved the solution S^* while still satisfying the constraint.

- Consider the first case. We have as for the proof of Proposition 20, that

$$\left(1 - e^{-1}\right) \frac{\sigma(S^*)}{\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*} \leq \frac{(1-\beta)\sigma(S_\ell) + \beta\sigma(S_{\ell+1})}{(1-\beta)\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c}^* + \beta\mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}^*},$$

where $\ell \in \{0, 1, \dots, |V|-1\}$ is such that $\mathbf{e}_{S_\ell}^\intercal \mathbf{c} \leq \mathbf{e}_{S^*}^\intercal \mathbf{c} \leq \mathbf{e}_{S_{\ell+1}}^\intercal \mathbf{c}$, and $\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^* = (1-\beta)\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c}^* + \beta \mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}^*$. In the case S_ℓ has a greater ratio than $S_{\ell+1}$, it is chosen by our algorithm and has the desired approximation. In the case $S_{\ell+1}$ has the better ratio, it is chosen if its cost is lower than b. If its cost is greater than b, then $\ell+1 = j$ and the algorithm chooses S_ℓ with some probability $1-\beta'$ and $S_{\ell+1}$ with probability β'. The goal is to show that the coefficient β' we use for $S_{\ell+1}$ is greater than β. This must be the case since $(1-\beta')\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c}^* + \beta' \mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}^* = b > \mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^* = (1-\beta)\mathbf{e}_{S_\ell\cup\{0\}}^\intercal \mathbf{c}^* + \beta \mathbf{e}_{S_{\ell+1}\cup\{0\}}^\intercal \mathbf{c}^*$.

- For the second case, we let S be the output of the Algorithm 1 considered by Wang, Yang, et al. (2020). We thus have from their Theorem 1 that

$$\left(1 - e^{-1}\right)\mathbb{E}[\sigma(S^*)] \leq \mathbb{E}[\sigma(S)].$$

Since $\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^*\right] = \mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right] = b$, we have

$$\left(1 - e^{-1}\right)\frac{\mathbb{E}[\sigma(S^*)]}{\mathbb{E}\left[\mathbf{e}_{S^*\cup\{0\}}^\intercal \mathbf{c}^*\right]} \leq \frac{\mathbb{E}[\sigma(S)]}{\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^*\right]}.$$

If the expected cost of the output S' of our algorithm is b, then both algorithms coincides and we have the desired result. Else, we have that S' maximizes the ratio over $\{S_0, \dots, S_j\}$, which contains the support of S (that is $\{S_{j-1}, S_j\}$), so the ratio evaluated at S' is greater than $\frac{\mathbb{E}[\sigma(S)]}{\mathbb{E}\left[\mathbf{e}_{S\cup\{0\}}^\intercal \mathbf{c}^*\right]}$, giving again the desired result.

□

6.4 Experiments and discussion

In this section, we present an experiment for Budgeted OIM. In Figure 6.1, we plot $\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \Delta(S_t)\right]$ with respect to the budget B used, running over up to $T = 10000$ rounds. This quantity is a good approximations to the true regret according to Proposition 19. Plotting the true regret would require to compute F_B^*, which is NP-Hard to do. We consider a subgraph of Facebook network (Leskovec and Krevl, 2014), with $|V| = 333$ and $|E| = 5038$, as in Wen, Kveton, Valko, et al. (2017). We take $\mathbf{w}^* \sim \mathcal{U}(0, 0.1)^{\otimes E}$ and take deterministic, known costs with $c_0^* = 1$, and $c_i^* = d_i / \max_{j\in V} d_j$. BOIM-CUCB$_+$ is the same approach as BOIM-CUCB$_5$ with bonus$(\,\cdot\,; \overline{\mathbf{w}}_{t-1})$ instead of bonus$_5$, ignoring that bonus$(\,\cdot\,; \overline{\mathbf{w}}_{t-1})$ is not submodular (it is only sub-additive).

We observed that in BOIM-CUCB$_1$, BOIM-CUCB$_4$, BOIM-CUCB$_5$, Condition 6.15 (with the correct bonus instead of bonus$_5$) always holds, meaning that those algorithms coincide with BOIM-CUCB *in practice*, and that the gain only appears through the analysis. We thus plot a single curve for these 4 algorithms in Figure 6.1. On the other hand, we observe only a slight gain of BOIM-CUCB+ compared to BOIM-CUCB.

Our experiments confirm that Fact 7 is less rough in practice, as we already anticipated. Indeed, our submodular bonuses are not tight enough to compete with BOIM-CUCB, although we gain in the analysis. The slight gain that we have for BOIM-CUCB+ suggests that the issue is not only about the tightness of a submodular upper bound, but rather about the tightness of Fact 7. This is supported by the following observation we made: for the Facebook subnetwork, for 1000 random draws

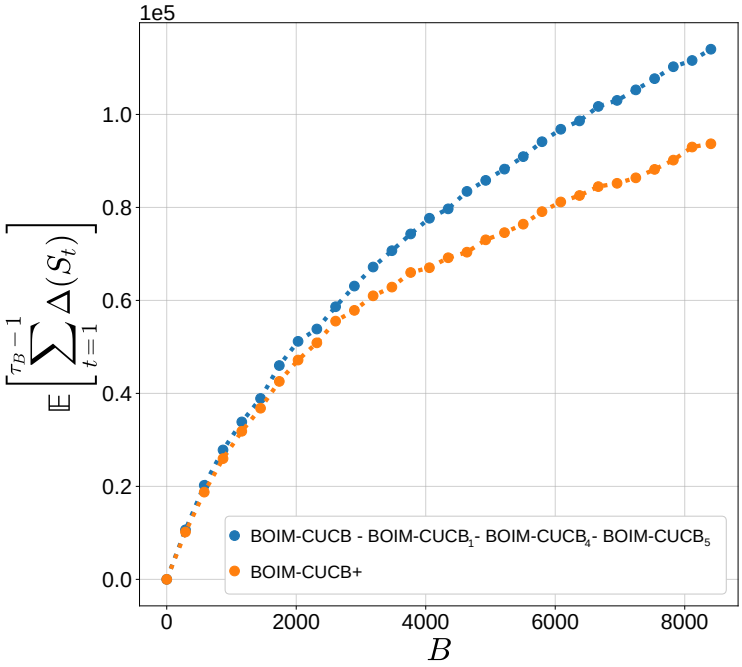

FIGURE 6.1: Regret curves with respect to the budget B (expectation computed by averaging over 10 independent simulations).

of a seed set and vector pairs in $[0, 0.1]^E$, the ratio of the RHS and the LHS in Fact 7 is each time greater than $0.4|V|$.

In the following, we conducted further experiments on a synthetic graph comparing BOIM-CUCB to BOIM-CUCB-REGULARIZED, which greedily maximizes the regularized spread $S \mapsto \sigma(S; \mathbf{w}_t) - \lambda \mathbf{e}_S \mathbf{c}_t$, where λ is a parameter to set. We observed that for an appropriate choice of λ, a performance similar to BOIM-CUCB can be obtained. The experiments were conducted on a complete 10 nodes graph, with known costs $c_0^* = 1$, and for all $i \in V$, $c_i^* \sim \mathcal{U}(0, 1)$. We also chose $\mathbf{w}^* \sim \mathcal{U}(0, 0.1)^{\otimes E}$, as previously. We compare the BOIM-CUCB algorithm to BOIM-CUCB-REGULARIZED, another algorithm that might challenge BOIM-CUCB in our setting. BOIM-CUCB-REGULARIZED is exactly as BOIM-CUCB except that the objective that is optimized is $S \mapsto \sigma(S; \mathbf{w}_t) - \lambda \mathbf{e}_S \mathbf{c}_t$, for λ being an input parameter to the algorithm. We can see that as for BOIM-CUCB, this algorithm have the willingness to maximize the function σ while minimizing the cost function. The fundamental difference is on the importance given to one or the other function, controlled by λ. We use a greedy maximization in BOIM-CUCB-REGULARIZED. A greedy optimization of the objective $S \mapsto \sigma(S; \mathbf{w}_t) - \lambda \mathbf{e}_S \mathbf{c}_t$ is a heuristic which, although not supported in theory, performs well in practice. We run experiments over up to $T = 10000$ rounds, on five different draws for $\mathbf{w}^*$ and $\mathbf{c}^*$, and 3 different values $\lambda = 2, 3, 4$. Results are shown in Figure 6.2. We observe that BOIM-CUCB is in general better than BOIM-CUCB-REGULARIZED. If the variable λ is properly chosen, performances similar to BOIM-CUCB can be obtained. This is not surprising since BOIM-CUCB aims (but only approximately) to select $S_t^* \in \arg\max_{S \subset V} \sigma(S; \mathbf{w}_t)/\mathbf{e}_{S \cup \{0\}}$. If $\lambda = \sigma(S_t^*; \mathbf{w}_t)/\mathbf{e}_{S_t^* \cup \{0\}} \mathbf{c}_t$, then we also have that BOIM-CUCB-REGULARIZED aims at choosing S_t^*, since one can notice that $S_t^* \in \arg\max_{S \subset V} \sigma(S; \mathbf{w}_t) - \lambda \mathbf{e}_S \mathbf{c}_t$.

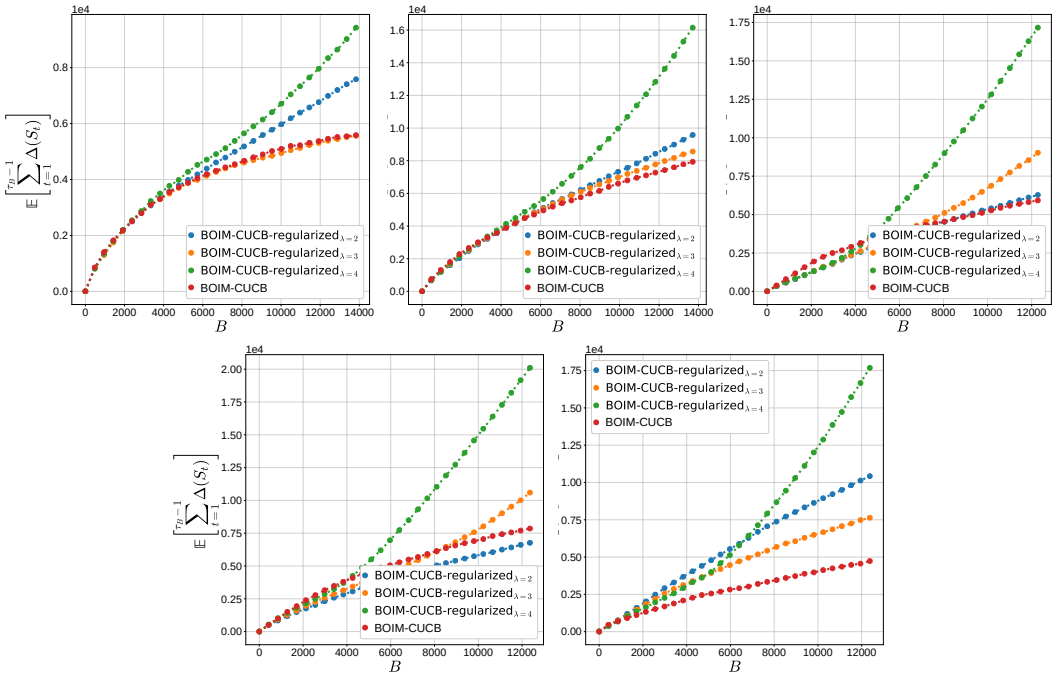

FIGURE 6.2: Regret curves on five different problem instances, with respect to the budget B (expectation computed by averaging over 10 independent simulations).

6.4.1 Discussion and Future work

We introduced a new Budgeted OIM problem, taking both the costs of influencers and fixed costs into account in the seed selection, instead of the usual cardinality constraint. This better represents the current challenges in viral marketing, since top influencers tend to be more and more costly. Our fixed cost can also be seen as the time that a round takes: A null fixed cost would mean that reloading the network to get a new independent instance is *free* and *instantaneous*.[7] Obviously, this is not realistic. We also provided an algorithm for Budgeted OIM under the IC model and the edge level semi-bandit feedback setting.

Interesting future directions of research would be to explore other kinds of feedback or diffusion models for Budgeted OIM. For practical scalability, it would also be good to investigate the incorporation of the linear generalization framework (Wen, Kveton, Valko, et al., 2017) into Budgeted OIM. Notice, this extension is *not* straightforward if we want to keep our tighter confidence region. More precisely, we believe that a *linear semi-bandit* approach that is aware of independence between edge observations should be developed (the linear generalization approach of Wen, Kveton, Valko, et al. (2017) treats each edge observation as *arbitrary correlated*).

In addition to this, exploring how the use of Fact 7 *in the Algorithm* might be avoided while still using confidence region given by Fact 8 would surely improve the algorithms. One possible way would be to use a Thompson Sampling (TS) approach (Wang and Chen, 2018; Perrault, Boursier, et al., 2020), where the prior takes into account the mutual independence of weights. However, Wang and Chen (2018) proved

[7] In this case, using $|S|$ rounds choosing each time a single different influencer $i \in S$ is better than choosing the whole S in a single round.

in their Theorem 2 that TS gives linear approximation regret for some special approximation algorithms. Thus, we would have to use some specific property of the GREEDY approximation algorithm we use.

6.5 Missing proofs

6.5.1 Proof of Proposition 19

Proof of Proposition 19. Let $\alpha = 1 - 1/e - \varepsilon$. In the proof, we shall consider several policies π one after the other. In each case, we will denote by S_t the seed selected by π at round t, and τ_B the random round where π has exhausted its budget.

Consider first the policy π that selects

$$S_t = S^* \in \arg\max_{S \subset V} \mathbb{E}[\sigma(S; \mathbf{W})] \mathbb{E}\left[\mathbf{e}_{S \cup \{0\}}^\intercal \mathbf{C}\right]^{-1}$$

at each round $t \geq 1$. We can write

$$\begin{aligned}
F_B^* + |V| &\geq F_B^* + \mathbb{E}[\sigma(S^*; \mathbf{W})] \\
&\geq F_B(\pi) + \mathbb{E}[\sigma(S^*; \mathbf{W})] && \text{definition of } F_B^* \\
&= \sum_{t \geq 1} \mathbb{E}[\sigma(S^*; \mathbf{W}_t) \mathbb{I}\{B_{t-1} \geq 0\}] \\
&= \sum_{t \geq 1} \mathbb{E}[\mathbb{E}[\sigma(S^*; \mathbf{W})] \mathbb{I}\{B_{t-1} \geq 0\}] && \text{conditioning on } \mathcal{H}_t \\
&= \lambda^* \sum_{t \geq 1} \mathbb{E}\left[\mathbb{E}\left[\mathbf{e}_{S^* \cup \{0\}}^\intercal \mathbf{C}\right] \mathbb{I}\{B_{t-1} \geq 0\}\right] \\
&= \lambda^* \sum_{t \geq 1} \mathbb{E}[(C_{0,t} + \mathbf{e}_{S^*}^\intercal \mathbf{C}_t) \mathbb{I}\{B_{t-1} \geq 0\}] && \text{conditioning on } \mathcal{H}_t \\
&\geq \lambda^* B && \text{definition of } \tau_B.
\end{aligned}$$

We can use the inequality

$$F_B^* + |V| \geq \lambda^* B \tag{6.19}$$

with any policy π in the following ways:

- First, we can bound the cost part in the cumulative gap:

$$\begin{aligned}
&\alpha \lambda^* \mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \mathbb{E}\left[\mathbf{e}_{S_t \cup \{0\}}^\intercal \mathbf{C}\right]\right] \\
&\leq \alpha \lambda^* \sum_{t \geq 1} \mathbb{E}\left[\mathbb{E}\left[\mathbf{e}_{S_t \cup \{0\}}^\intercal \mathbf{C}\right] \mathbb{I}\{B_{t-1} \geq 0\}\right] && B_t \geq 0 \Rightarrow B_{t-1} \geq 0 \\
&= \alpha \lambda^* \sum_{t \geq 1} \mathbb{E}[(\mathbf{e}_{S_t}^\intercal \mathbf{C}_t + C_{0,t}) \mathbb{I}\{B_{t-1} \geq 0\}] && \text{conditioning on } \mathcal{H}_t \\
&\leq \alpha \lambda^* B + \alpha \lambda^* (1 + |V|) && \text{definition of } \tau_B, \\
&\leq \alpha F_B^* + \alpha |V| + \alpha \lambda^* (1 + |V|) && \text{inequality (6.19)}.
\end{aligned}$$

- Next, we can bound the reward part:

$$\mathbb{E}\left[\sum_{t=1}^{\tau_B - 1} \mathbb{E}[\sigma(S_t; \mathbf{W})]\right]$$

$$
\begin{aligned}
&\geq \mathbb{E}\left[\sum_{t=1}^{\tau_B} \mathbb{E}[\sigma(S_t;\mathbf{W})]\right] - |V| \\
&= \sum_{t\geq 1} \mathbb{E}[\mathbb{E}[\sigma(S_t;\mathbf{W})]\mathbb{I}\{B_{t-1}\geq 0\}] - |V| \\
&= \sum_{t\geq 1} \mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_{t-1}\geq 0\}] - |V| && \text{conditioning on } \mathcal{H}_t \\
&\geq \sum_{t\geq 1} \mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_t\geq 0\}] - |V| && B_t\geq 0 \Rightarrow B_{t-1}\geq 0 \\
&= F_B(\pi) - |V|.
\end{aligned}
$$

Adding these two inequalities, we get the following upper bound on the cumulative gap:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\Delta(S_t)\right] &= \alpha\lambda^*\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}\left[\mathbf{e}_{S_t\cup\{0\}}^\intercal\mathbf{C}\right]\right] - \mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}[\sigma(S_t;\mathbf{W})]\right] \\
&\leq R_{B,\varepsilon}(\pi) + (\alpha+1)|V| + \alpha\lambda^*(1+|V|).
\end{aligned}
$$

In the same way, we can derive a lower bound on $\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\Delta(S_t)\right]$, considering first the policy π such that $F_B(\pi) = F_B^*$:

$$
\begin{aligned}
F_B^* &= \sum_{t\geq 1}\mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_t\geq 0\}] \\
&\leq \sum_{t\geq 1}\mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_{t-1}\geq 0\}] && B_t\geq 0\Rightarrow B_{t-1}\geq 0 \\
&= \sum_{t\geq 1}\mathbb{E}[\mathbb{E}[\sigma(S_t;\mathbf{W})]\mathbb{I}\{B_{t-1}\geq 0\}] && \text{conditioning on } \mathcal{H}_t \\
&\leq \sum_{t\geq 1}\mathbb{E}\left[\lambda^*\mathbb{E}\left[\mathbf{e}_{S_t\cup\{0\}}^\intercal\mathbf{C}\right]\mathbb{I}\{B_{t-1}\geq 0\}\right] && \text{definition of } \lambda^* \\
&= \lambda^*\mathbb{E}\left[\sum_{t=1}^{\tau_B}(C_{0,t}+\mathbf{e}_{S_t}^\intercal\mathbf{C}_t)\right] && \text{conditioning on } \mathcal{H}_t \\
&\leq \lambda^*(B+1+|V|) && \text{definition of } \tau_B.
\end{aligned}
$$

i.e.,

$$
F_B^* - \lambda^*(1+|V|) \leq \lambda^* B. \tag{6.20}
$$

Considering any policy π:

$$
\begin{aligned}
&\alpha\lambda^*\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}\left[\mathbf{e}_{S_t\cup\{0\}}^\intercal\mathbf{C}\right]\right] \\
&\geq \alpha\lambda^*\sum_{t\geq 1}\mathbb{E}\left[\mathbb{E}\left[\mathbf{e}_{S_t\cup\{0\}}^\intercal\mathbf{C}\right]\mathbb{I}\{B_{t-1}\geq 1+|V|\}\right] && B_{t-1}\geq 1+|V|\Rightarrow B_t\geq 0 \\
&= \alpha\lambda^*\sum_{t\geq 1}\mathbb{E}[(\mathbf{e}_{S_t}^\intercal\mathbf{C}_t + C_{0,t})\mathbb{I}\{B_{t-1}\geq 1+|V|\}] && \text{conditioning on } \mathcal{H}_t \\
&\geq \alpha\lambda^* B - \alpha\lambda^*(1+|V|) && \text{definition of } \tau_B. \\
&\geq \alpha F_B^* - 2\alpha\lambda^*(1+|V|) && \text{inequality (6.20)},
\end{aligned}
$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}[\sigma(S_t;\mathbf{W})]\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\tau_B}\mathbb{E}[\sigma(S_t;\mathbf{W})]\right]\\ &= \sum_{t\geq 1}\mathbb{E}[\mathbb{E}[\sigma(S_t;\mathbf{W})]\mathbb{I}\{B_{t-1}\geq 0\}]\\ &= \sum_{t\geq 1}\mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_{t-1}\geq 0\}] \qquad \text{conditioning on } \mathcal{H}_t\\ &\leq \sum_{t\geq 1}\mathbb{E}[\sigma(S_t;\mathbf{W}_t)\mathbb{I}\{B_t\geq 0\}] + |V|\\ &= F_B(\pi) + |V|.\end{aligned}$$

Again adding these two inequalities, we get the desired lower bound:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\Delta(S_t)\right] &= \alpha\lambda^*\mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}\left[\mathbf{e}_{S_t\cup\{0\}}^\intercal\mathbf{C}\right]\right] - \mathbb{E}\left[\sum_{t=1}^{\tau_B-1}\mathbb{E}[\sigma(S_t;\mathbf{W})]\right]\\ &\geq R_{B,\varepsilon}(\pi) - 2\alpha\lambda^*(1+|V|) - |V|.\end{aligned}$$

□

6.5.2 Proof of Proposition 21

Proof of Proposition 21. First, notice that we trivially have

$$\mathbb{P}\left[\mathfrak{P}_t \text{ and } p_i(\{j\};\mathbf{w}^*) \leq \frac{8\delta(t)}{N_{\ominus j,t-1}} \text{ and } \frac{\delta(t)p_i(\{j\};\mathbf{w}^*)}{N_{\oplus i,t-1}} > \frac{8\delta(t)}{N_{\ominus j,t-1}}\right] = 0.$$

Thus, let's prove that

$$\mathbb{P}\left[\mathfrak{P}_t \text{ and } p_i(\{j\};\mathbf{w}^*) > \frac{8\delta(t)}{N_{\ominus j,t-1}} \text{ and } \frac{\delta(t)p_i(\{j\};\mathbf{w}^*)}{N_{\oplus i,t-1}} > \frac{8\delta(t)}{N_{\ominus j,t-1}}\right] \leq 1/t^2.$$

We define another counter for $(i,j)\in V^2$ as follows:

$$N_{i,j,t-1} \triangleq \sum_{t'=1}^{t-1}\mathbb{I}\left\{j\in S_{t'},\ \{j\} \overset{\mathbf{W}_{t'}}{\rightsquigarrow} i\right\}.$$

Note that we have $N_{i,j,t-1}\leq (N_{\oplus i,t-1}\wedge N_{\ominus j,t-1})$. We can thus remove $\mathfrak{P}_t$ and replace $N_{\oplus i,t-1}$ by $N_{i,j,t-1}$, since this can only increases the probability. By an union bound we have,

$$\begin{aligned}&\mathbb{P}\left[p_i(\{j\};\mathbf{w}^*) > \frac{8\delta(t)}{N_{\ominus j,t-1}} \text{ and } \frac{p_i(\{j\};\mathbf{w}^*)}{N_{i,j,t-1}} > \frac{8}{N_{\ominus j,t-1}}\right]\\ &\leq \sum_{t'>\frac{8\delta(t)}{p_i(\{j\};\mathbf{w}^*)}}\mathbb{P}\left[N_{\ominus j,t-1}=t',\ \frac{t'p_i(\{j\};\mathbf{w}^*)}{8} > N_{i,j,t-1}\right].\end{aligned}$$

Since the random variables $\mathbb{I}\left\{\{j\} \overset{\mathbf{w}_{t''}}{\rightsquigarrow} i\right\}$ are bernouillies of mean $p_i(\{j\}; \mathbf{w}^*)$, we can apply the Fact 9 to get

$$\mathbb{P}\left[N_{\ominus j,t-1} = t',\ \frac{t' p_i(\{j\}; \mathbf{w}^*)}{8} > N_{i,j,t-1}\right] \leq \exp\left(-(7/8)^2 8\delta(t)/2\right) < 1/t^3.$$

By taking t' over $\{0, \ldots, t-1\}$, the proposition holds.

Fact 9 (Multiplicative Chernoff Bound Mitzenmacher and Upfal (2017))**.** *Consider $X_1, \ldots, X_t$ be Bernoulli random variables, of parameter μ, then for $Y = X_1 + \cdots + X_t$, we have with $\delta \in (0, 1)$,*

$$\mathbb{P}[Y \leq (1-\delta)t\mu] \leq e^{-\delta^2 t\mu/2}.$$

□

Chapter 7

Covariance-Adapting Policy

This chapter is based on our paper Perrault, Perchet, and Valko (2020). Its aim is to offer a more practical, and more accurate solution for building elliptical confidence region. Indeed, the ones we have seen so far are either based on the mutual independence of outcomes, or due to the unrealistic knowledge of a sub-Gaussianity matrix $\mathbf{\Gamma}$. We alleviate this issue by instead considering a new general family of *sub-exponential* distributions, which contains bounded and Gaussian ones. We prove a new lower bound on the regret on this family, that is parameterized by the *unknown* covariance matrix, a tighter quantity than the sub-Gaussianity matrix. We then construct an algorithm that uses covariance estimates, and provide a tight asymptotic analysis of the regret. Finally, we apply and extend our results to the family of sparse outcomes, which has applications in many recommender systems.

7.1 An alternative to sub-Gaussian outcomes

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

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

In all this chapter, we consider a linear reward function. To recall, at each round t, the agent chooses some action $A_t \in \mathcal{A}$, receives the *total* reward associated to the selected actions A_t, assumed to be $\sum_{i \in A_t} X_{i,t}$, and observes the outcome of each base arm of A_t, i.e., the vector $(X_{i,t}\mathbb{I}\{i \in A_t\})_{i \in [n]}$. The action space $\mathcal{A}$ depends on the combinatorial problem at hand. For example, actions in $\mathcal{A}$ could be a path from an origin to a destination in a network (György et al., 2007; Talebi, Zou, et al., 2013) or a subset of items to recommend to a customer (Wang, Ouyang, et al., 1997). Many other examples and applications are given by Cesa-Bianchi and Lugosi

(2012). As we already mentioned, in this setting, the whole joint distribution of the vector of outcomes is relevant, contrary to standard bandit problems where only the n marginals are sufficient to characterize the difficulty of the instance. If we define $\mathbf{X} \triangleq (X_1, \dots, X_n)$, the objective is to design a learning algorithm adaptive to the distribution $\mathbb{P}_{\mathbf{X}}$. This is more challenging than in standard bandits, where adaptivity is only with respect to $\otimes_{i\in[n]}\mathbb{P}_{X_i}$.

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

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^n,\ \mathbb{E}\Big[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\Big] \leq e^{\boldsymbol{\lambda}^\intercal \mathbf{C}\boldsymbol{\lambda}/2}. \tag{7.1}$$

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

$$\frac{\log T}{\Delta} \sum_{i\in[n]} \Gamma_{ii}\Big((1-\gamma)\log^2(m) + \gamma m\Big), \tag{7.2}$$

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

Our main objective is to alleviate this issue, and to strive to obtain fast rates for combinatorial semi-bandits, as Degenne and Perchet (2016b) in the case where there is a favorable covariance structure, but *without knowing* it beforehand. Therefore, algorithms should be able to capture the covariance structure given by $\boldsymbol{\Gamma}$ from the data processed and adapt to it. We actually go further by asking whether the matrix $\boldsymbol{\Gamma}$ is the relevant parameter to characterize the difficulty of a problem. We argue that the covariance matrix $\boldsymbol{\Sigma}^* \triangleq \mathbb{E}\Big[(\mathbf{X}-\boldsymbol{\mu}^*)(\mathbf{X}-\boldsymbol{\mu}^*)^\intercal\Big]$ is more pertinent, as it allows to better differentiate complex problems from the easy ones. One can indeed already argue in favor of a $\boldsymbol{\Sigma}^*$ dependence rather than a $\boldsymbol{\Gamma}$ one, based on the relation $\boldsymbol{\Sigma}^* \preceq_+ \boldsymbol{\Gamma}$. Indeed, this results from the fact that $\boldsymbol{\Sigma}^* \preceq \mathbf{C}$, which can be proved as follows: Fix $\mathbf{x} \in \mathbb{R}^n$. For any $\lambda \in \mathbb{R}$, $\mathbb{E}\Big[e^{\lambda\mathbf{x}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\Big] \leq e^{\lambda^2\mathbf{x}^\intercal\mathbf{C}\mathbf{x}/2}$. The second order Taylor expansion in λ gives

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

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

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

(*i*) The matrix $\mathbf{\Gamma}$ needs to be known. This requires specific knowledge about the outcome structure, which is often not precise, as it is usually only known that outcomes are bounded, or at most that there exists some constant κ such that $\kappa^2 \geq C_{ii}$ for all $i \in [n]$. The latter is equivalent[1] to $\Gamma_{ij} = \kappa^2$ for all $i, j \in [n]$ and corresponds to the worst case correlation between outcomes ($\gamma = 1$) in the regret bound (5.4).

(*ii*) The value γ can be 1, even when outcomes are only weakly correlated: For instance, if n is even, $\mathbf{\Gamma}$ can be a block-diagonal matrix with $n/2$ blocks of size 2×2 containing only ones. This scenario can actually occur in many examples; we provide two types below:

- Arms are nodes on a given graph, with some small communities on which outcome tends to be constant (Cesa-Bianchi, Gentile, and Zappella, 2013; Valko et al., 2014; Gentile, Li, and Zappella, 2014; Valko, 2016).
- Arms are market-basket-like items, with some highly correlated pairs of items (e.g., people buying from category "books" tend to also buy from category "CDs", Zhang and Feigenbaum, 2006; He, Xu, and Deng, 2006).

(*iii*) The value Γ_{ii} can be high, even for low-variance outcomes, while intuitively, low variance outcomes should be easy to work with. For example, if $\mathbf{X}$ is a binary 1-sparse random variable — as in some recommender systems, where a single item is desired by the user — then $X_i \sim \text{Bernoulli}(\mu_i^*)$ with $\sum_{i=1}^n \mu_i^* = 1$, and $\Gamma_{ii} \geq C_{ii} \geq (\mu_i^* - 1/2)/(\log(\mu_i^*) - \log(1 - \mu_i^*))$ (and this is tight, see, e.g., Buldygin and Moskvichova, 2013). For μ_i^* of order $1/n$, Γ_{ii} is thus at least of order $1/(2 \log n)$ for n large, whereas $\mathbb{V}(X_i)$ is of order $1/n$.

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

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

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

Our main contribution is in the analysis of this approach, that is not based on the usual *Laplace's method*, which works in the sub-Gaussian framework, but does not handle well our sub-exponential-type assumption. Thus, our analysis is rather based on a *covering-argument* (Magureanu, Combes, and Proutiere, 2014). An important part of our proof is based on the transformation of the axis-unaligned ellipsoidal confidence region associated to a given action $A \in \mathcal{A}$ into an axis-aligned region,

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

using the following relation $\left(\Sigma^*_{ij}\right)_{ij\in A} \preceq_+ \operatorname{diag}\left(\sum_{j\in A} 0 \vee \Sigma^*_{ij}\right)_{i\in A}$. This allows us to conduct the same type of proof than for the independent outcome case (where confidence regions are always axis-aligned), but with a Bernstein-type analysis.[2]

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

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

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

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

Sparse bandits Independently to combinatorial bandits, there exists a different setting actually dealing with correlated outcomes in online learning known as *sparse bandits* (Kwon, Perchet, and Vernade, 2017; Kwon and Perchet, 2015; Bubeck, Cohen, and Li, 2017; Abbasi-Yadkori, Pal, and Szepesvari, 2012; Carpentier and Munos, 2012; Gerchinovitz, 2013). The overall idea is to introduce by now a standard sparsity assumption (some parameter vector has only s out of its n components that are non zero) into sequential decision making. As usual, the objective is to replace the

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

linear/polynomial dependence in the dimension n by a linear/polynomial dependence in s. Quite interestingly, the sparsity assumption has been studied in two different directions. The fist one assumes that the vector $\boldsymbol{\mu}^*$ is s-sparse, typically in (linear) stochastic bandits (Kwon, Perchet, and Vernade, 2017; Abbasi-Yadkori, Pal, and Szepesvari, 2012; Carpentier and Munos, 2012; Gerchinovitz, 2013). The second one assumes that the realized vector $\mathbf{X}_t$ is s-sparse, usually in adversarial bandits (Kwon and Perchet, 2015; Bubeck, Cohen, and Li, 2017).

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

7.1.1 A multivariate sub-exponential distribution

Let us recall that the goal for the agent is to minimize the regret, that is defined as follows, with $A^* \in \arg\max_{A \in \mathcal{A}} \mathbf{e}_A^\intercal \boldsymbol{\mu}^*$,

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

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

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

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

Assumption 3 ($\|\cdot\|_1$-sub-exponential distribution)**.** *$\forall \boldsymbol{\lambda} \in \mathbb{R}^n$ such that $\|\boldsymbol{\lambda}\|_1 \leq 1/(2\kappa)$, we have $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal (\mathbf{X} - \boldsymbol{\mu}^*)}\right] \leq e^{\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^* \boldsymbol{\lambda}}$, where $\boldsymbol{\Sigma}^* \triangleq \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}^*)(\mathbf{X} - \boldsymbol{\mu}^*)^\intercal\right]$ is the covariance matrix of $\mathbf{X}$.*

Importantly, the agent does not know the covariance matrix $\boldsymbol{\Sigma}^*$. Remark that Assumption 3 trivially holds for $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\forall \boldsymbol{\lambda} \in \mathbb{R}^n$, $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal (\mathbf{X} - \boldsymbol{\mu}^*)}\right] = e^{\boldsymbol{\lambda}^\intercal \boldsymbol{\Sigma}^* \boldsymbol{\lambda} / 2}$. The following proposition states that it also holds for bounded outcomes.

Proposition 24. *If $\|\mathbf{X}\|_\infty \leq \kappa$, then both Assumption 2 and 3 hold.*

Proof. Assumption 2 is a direct consequence of Hoeffding's Lemma. For Assumption 3, we have $\|\mathbf{X} - \boldsymbol{\mu}^*\|_\infty \leq 2\kappa$. For $\|\boldsymbol{\lambda}\|_1 \leq 1/(2\kappa)$, we have:

$$\begin{aligned}
&\log \mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal (\mathbf{X} - \boldsymbol{\mu}^*)}\right] \\
&= \log\left(1 + \sum_{k \geq 2} \mathbb{E}\left[\frac{(\boldsymbol{\lambda}^\intercal (\mathbf{X} - \boldsymbol{\mu}^*))^k}{k!}\right]\right)
\end{aligned}$$

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

□

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

7.2 Covariance-dependent regret (lower) bound

7.2.1 Lower bound

We start by proving in Theorem 31 a new gap-dependent lower bound on R_T, valid for any covariance matrix $\boldsymbol{\Sigma}^* \succeq 0$, for some $\mathbb{P}_{\mathbf{X}}$ satisfying Assumptions 2 and 3, some action space $\mathcal{A}$, and for any consistent algorithm (Lai and Robbins, 1985), for which the regret on any problem verifies $R_T = o(T^a)$ as $T\to\infty$, for all $a>0$. This lower bound demonstrates the link between $\boldsymbol{\Sigma}^*$ and the difficulty of the problem. It also indicates, in anticipation, that we have to examine a subclass of action sets to hope to improve the upper bound we will provide in Theorem 32.

Theorem 31. *For any $n, m \in \mathbb{N}^*$ such that $n/m \geq 2$ is an integer, any $n\times n$ matrix $\boldsymbol{\Sigma}^* \succeq 0$, any $\Delta>0$, and any consistent policy, there exists an instance with n arms — characterized by some action space $\mathcal{A}$, with $m = \max\{|A|,\ A\in\mathcal{A}\}$, some outcome distribution $\mathbb{P}_{\mathbf{X}}$ satisfying Assumptions 2 and 3 with all gaps equal to Δ and covariance matrix $\boldsymbol{\Sigma}^*$ — on which the regret satisfies*

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

The proof considers $\mathcal{A}$ containing n/m disjoint actions $A_1,\dots,A_{n/m}$ composed of m arms, with $A_k = \{(k-1)m+1,\dots,km\}$, and $\mathbf{X}\sim\mathcal{N}(-\Delta/m(\mathbb{I}\{i\notin A_1\})_i, \boldsymbol{\Sigma}^*)$. The idea is to make a reduction to some standard bandit problems with n/m arms, and to compute the number of rounds t needed to distinguish between A_k and A_1. Roughly speaking, t is at least equal to the inverse of the KL between outcome distributions of A_k and its centered version, and in the case of Gaussian distributions, we get $t \geq 2\mathbb{V}(\sum_{i\in A_k} X_i)/\Delta^2 = 2\mathbf{e}_{A_k}^\intercal\boldsymbol{\Sigma}^*\mathbf{e}_{A_k}/\Delta^2$. It is not surprising that the variance appears, since this can be seen as a measure of the uncertainty we have in our samples: The higher the variance, the harder the estimation, and therefore the higher the round t must be. Notice that Theorem 31 is a refinement of Theorem 1 from Degenne and Perchet (2016b), in which they consider the same action space $\mathcal{A}$ but a specific choice for the matrix $\boldsymbol{\Sigma}^*$: it is a block-diagonal matrix with n/m blocks, where each block (corresponding to an action A) is equal to $\sigma^2((1-\gamma)\mathrm{diag}(\mathbf{e}_A) + \gamma\mathbf{e}_A\mathbf{e}_A^\intercal)$, i.e., they take the worst case correlation under the controls given by σ^2 and γ, and knowing

that the problem given by $\mathcal{A}$ is agnostic to the correlations between the arms of two different blocks.

Proof of Theorem 31. Consider $\mathcal{A}$ containing n/m disjoint actions $A_1, \dots, A_{n/m}$ composed of m arms, with $A_k = \{(k-1)m+1, \dots, km\}$, and

$$\mathbf{X} \sim \mathcal{N}(-\Delta/m(\mathbb{I}\{i \notin A_1\})_i, \mathbf{\Sigma}^*).$$

This problem reduces to a standard bandit problem with n/m arms. We use a result from Burnetas and Katehakis (1996), a generalization of Lai and Robbins (1985), that states that

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

As we can write

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

it holds that

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

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

In the next subsection, we describe our algorithm ESCB-C (Algorithm 11) and provide an upper bound on its regret in Theorem 32, where the expression

$$\max_{A\in\mathcal{A}, i\in A} \sum_{j\in A} 0 \vee \Sigma^*_{ij}$$

appears. Notice that this is very close to the expression given in Theorem 31. In fact, both expressions coincide when $\mathbf{\Sigma}^*$ has only non-negative entries.

7.2.2 Main algorithm and the guarantees

In this section, we present an algorithm for the setting introduced in Section 4.1.1. The method is stated as Algorithm 11. To find the action with the highest mean, the agent estimates the mean μ_i^* of every arm i with their corresponding *empirical averages* defined as $\overline{\mu}_{i,t-1} \triangleq \sum_{u\in[t-1]} \frac{\mathbb{I}\{i\in A_u\} X_{i,u}}{N_{i,t-1}}$, for $t \geq 1$, where $N_{i,t-1} \triangleq \sum_{u\in[t-1]} \mathbb{I}\{i \in A_u\}$ is the number of time arm i have been drawn for the first $t-1$ rounds. As mentioned above, the agent also estimates the covariance $\Sigma^*_{ij} = \mathbb{E}[X_i X_j] - \mu_i^* \mu_j^*$ of each pair $(i,j) \in [n]^2$. This will be done with the following estimate

$$\overline{\Sigma}_{ij,t-1} \triangleq \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j \in A_u\}\Big(X_{i,u} - \overline{\mu}_{i,t-1}\Big)\Big(X_{j,u} - \overline{\mu}_{j,t-1}\Big)}{N_{ij,t-1}}$$

$$= \sum_{u\in[t-1]} \frac{\mathbb{I}\{i,j\in A_u\}\Big(X_{i,u}X_{j,u} - \overline{\mu}_{i,t-1}X_{j,u} - \overline{\mu}_{j,t-1}X_{i,u}\Big)}{N_{ij,t-1}} + \overline{\mu}_{i,t-1}\overline{\mu}_{j,t-1},$$

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

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

Using concentration inequalities, we get confidence intervals for the above estimates. We are then able to use an upper-confidence-bound strategy (Auer, Cesa-Bianchi, and Fischer, 2002). More precisely, we first build the upper confidence bound on Σ^*_{ij} using the fact that $X_i \cdot X_j$ is a *sub-exponential* random variable, since both X_i and X_j are sub-Gaussian by virtue of Assumption 7.1. The result is stated in the following proposition.

Proposition 25. *With probability* $1-10t^{-2}$*, we have*

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

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

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

$$\overline{\Sigma}_{ij,t-1} = \widetilde{\Sigma}_{ij,t-1} + \Big(\mu^*_i - \overline{\mu}_{i,t-1}\Big)\Big(\widetilde{\mu}_{j,t-1} - \overline{\mu}_{j,t-1}\Big) + \Big(\mu^*_j - \overline{\mu}_{j,t-1}\Big)(\widetilde{\mu}_{i,t-1} - \mu^*_i).$$

We now state Lemma 10 giving sub-exponential parameters for a product of sub-Gaussian random variables. A proof comes from Honorio and Jaakkola (2014).

Lemma 10. *If* Y, Z *are* 1*-sub-Gaussian random variables, then* $\forall|\lambda| \leq 1/8$*,*

$$\mathbb{E}\Big[e^{\lambda(YZ-\mathbb{E}[YZ])}\Big] \leq e^{64\lambda^2}.$$

We apply Lemma 10 with a Chernoff argument and an union bound (to avoid the randomness of counters) in order to get the following Bernstein inequality

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

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

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

which is enough to conclude the proof. Notice that for the second inequality above, we take the union bound for two counters. When they are not random,

$$N_{ij,t-1}\Big(\widetilde{\mu}_{j,t-1} - \overline{\mu}_{j,t-1}\Big),$$

that is equal to

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

is a sum of $N_{j,t-1}$ independent random variables, $N_{ij,t-1}$ of which are $\left(1-\frac{N_{ij,t-1}}{N_{j,t-1}}\right)^2$-sub-Gaussian and the remaining ones are $\frac{N_{ij,t-1}^2}{N_{j,t-1}^2}$-sub-Gaussian. Therefore, this random variable is $N_{ij,t-1}\Big(1-\frac{N_{ij,t-1}}{N_{j,t-1}}\Big)$-sub-Gaussian, and in particular $N_{ij,t-1}$ -sub-Gaussian. □

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

$$\mathcal{C}_t(A) \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{\boldsymbol{\xi}\in\mathbb{R}^n, \sum_{i\in A}\frac{N_{i,t-1}\xi_i^2}{|A||\xi_i| + \sum_{j\in A} 0\vee\Sigma_{ij,t}} \leq 8(\log t + \log\log t) + 4em\right\}. \tag{7.3}$$

The intuition behind this confidence region is similar to the one for empirical Bernstein confidence intervals, but the term $\sum_{j\in A} 0\vee\Sigma_{ij,t}$ in the denominator replaces the usual empirical variance. To compare our confidence region with the one of Degenne and Perchet, 2016b, notice first that their algorithm uses the matrix $\boldsymbol{\Gamma}$ to build a confidence ellipsoid. They provide an analysis for this confidence ellipsoid using the *Laplace's method* and the matrix relation $\mathbf{C}\preceq_+\boldsymbol{\Gamma}$. In contrast, our confidence region is based on the covariance matrix $\boldsymbol{\Sigma}^*$. Our analysis is also different, as we use a *covering-argument* analysis. This is because the covariance estimation and Assumption 3 are both hard to handle with Laplace's method, that is more appropriate for sub-Gaussian random variables. Indeed, all calculations can be explicit and it is easy to construct a *conjugate prior*. This is *not the case* for sub-exponential random variables. Covering arguments are much more easier to use together with a diagonal matrix, so axis-aligned confidence region are desirable. We use an *axis-realignment technique* based on the matrix relation $\left(\Sigma_{ij}^*\right)_{ij\in A}\preceq_+ \operatorname{diag}\left(\sum_{j\in A} 0\vee\Sigma_{ij}^*\right)_{i\in A}$. The upside is to avoid dealing with off-diagonal terms by transforming them into diagonal ones. From all these previous observations, we can say that the confidence ellipsoid of Degenne and Perchet, 2016b is tighter as it does not require any axis realignment; however, not

Algorithm 11 ESCB-C *(Efficient Sampling for Combinatorial Bandits with Covariance estimate)*

Initialization:
Play $A_1 = [n]$, or at least a sequence $A_1, A_2, \dots,$ (no more than $n(n-1)/2$) such that for any $i, j \in [n]$, one of these A_t's contains $\{i, j\}$. We thus have $N_{ij,t-1} \geq 1$ for all $i, j \in [n]$.
For all subsequent rounds t:
Solve the following bilinear program to get A_t, with $\mathcal{C}_t(A)$ defined by (7.3), and play A_t,

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

only the matrix $\boldsymbol{\Gamma}$ is generally looser than $\boldsymbol{\Sigma}^*$ but also axis realignment does not alter the analysis, so that our new approach outperforms theirs in terms of asymptotic regret upper bound.

As common in bandits, the major challenge in the analysis is to prove that with high probability, $\boldsymbol{\mu}^* \in \mathcal{C}_t(A)$ for any action $A \in \mathcal{A}$. The covering argument together with the conversion from an axis-unaligned confidence region into an axis-aligned confidence region allows us to achieve this result (see Lemma 12). Therefore, an optimistic estimate $\boldsymbol{\mu}_t$ of the true mean $\boldsymbol{\mu}^*$ can be found using an upper-confidence-bound approach: if $A_t, \boldsymbol{\mu}_t$ are defined as in Algorithm 11, then, since $\boldsymbol{\mu}^* \in \mathcal{C}_t(A^*)$, we have

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

The regret bound for ESCB-C is stated in Theorem 32, and proven in subsection 7.5.2.

Theorem 32. *Assume that the outcome distribution $\mathbb{P}_\mathbf{X}$ satisfies Assumptions 7.1 and 3, and define $\Delta \triangleq \min_{A \in \mathcal{A},\ \Delta(A) > 0} \Delta(A)$, $\Delta_{\max} \triangleq \max_{A \in \mathcal{A},\ \Delta(A) > 0} \Delta(A)$. If Δ is small enough, i.e., there exists a universal constant c such that*

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

then the regret of Algorithm 11 is upper bounded as

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

where c' is a universal constant.

Notice that the bound in Theorem 32 is tight, up to a poly-logarithmic factor in m, with respect to the lower bound in Theorem 31, in the case where $\boldsymbol{\Sigma}^*$ has non-negative entries. Moreover, we focus on the asymptotic behavior of the regret (w.r.t. T) when Δ is small, i.e., when the problem becomes very difficult. While the quantity $c \log^2(m+1) \sum_{i \in [n]} \max_{A \in \mathcal{A}, i \in A} \sum_{j \in A} 0 \vee \Sigma^*_{ij} \log(T)/\Delta$ presented in Theorem 32 highlights the main dependence on both Δ and T, we prove a more precise non-asymptotic upper bound in (7.12), which holds for all $\Delta > 0$. Indeed, as for UCB-V, the errors from estimating $\boldsymbol{\Sigma}^*$ generate an extra term in the upper bound. However, since these errors are multiplied with estimation errors on the means, their impact is of second order. In particular, for Δ small enough, this extra term becomes negligible compared to the main term. Therefore, the term from covariance estimation errors is *not* present

in Theorem 32, but appears when Δ is far from 0. Finally, remark that when the covariance $\mathbf{\Sigma}^*$ is known, then one can consider the confidence region where $\Sigma_{ij,t}$ is replaced by Σ^*_{ij}. This avoids covariance estimation errors, and gives the upper bound of Theorem 32 when $\Delta + \Delta \log(\Delta_{\max}/\Delta)$ is smaller than $\sum_{i\in[n]} \max_{A\in\mathcal{A}, i\in A} \sum_{j\in A} 0 \vee \frac{\Sigma^*_{ij}}{n \cdot m}$.

Remark 15. *Considering the intersection of the region from Algorithm 11 with the one of* CUCB-V*, we can replace* $\log^2(m+1) \sum_{j\in A} 0 \vee \Sigma^*_{ij}$ *by*

$$(m\Sigma^*_{ii}) \wedge \left(\log^2(m+1) \sum_{j\in A} 0 \vee \Sigma^*_{ij} \right)$$

in Theorem 32.

7.3 Sparse outcomes

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

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

i.e., the number of nonzero components of $\mathbf{X}$ is smaller than s, where s is a fixed known parameter.[3] Importantly, the set of components which are nonzero is *not fixed nor known*, and may change over time. It should be noted, however, that there is a significant difference between the stochastic and the adversarial cases: In the later, the set of components which are nonzero change arbitrarily over time, whereas in the former, this set is sampled i.i.d. Notice, this sparse stochastic setting is different than the usual stochastic sparse bandit, where $\boldsymbol{\mu}^*$ is assumed to be sparse; see e.g., Kwon, Perchet, and Vernade (2017) for the classical MAB setting, and Abbasi-Yadkori, Pal, and Szepesvari (2012) and Carpentier and Munos (2012) for the linear bandit setting. For simplicity, we further assume that $\|\mathbf{X}\|_\infty \leq 1$. As we already saw in Proposition 24, this implies Assumption 2 and 3. The difficulty of this setting is that both the approach of Degenne and Perchet (2016b) and standard methods such as CUCB-V would not reach the lower bound for the regime $s \leq m$, as we will see. The reason is that a correlation exists between the components, because of sparsity, and must be taken into account.

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

Other examples involve settings where outcomes are usually zero except on very rare occasions: In the online routing, the packets are sent in a network and are lost if a server of that network has a failure. Because of failsafe procedures, failures

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

are *desynchronized* and typically only one (or at most a few) of them can happen simultaneously. In all of these examples, the decision maker has some combinatorial problem to solve: select an admissible path, select a *diverse* bundle of object/ads to display, etc., and only a few of the base items will generate non-zero outcome.

7.3.1 Lower bound

To start our study of sparse outcomes, we state a new lower bound in Theorem 33, that is valid for the setting described above. This lower bound is built on the same ideas as Theorem 31, with a notable variation: when reducing to an MAB problem, we do not obtain the necessary conditions for the application of Lai and Robbins (1985), because of the linear dependence between the μ_i^*'s. Thus, we use instead the the lower bound from Graves and Lai (1997). More precisely, we consider the same action space $\mathcal{A}$, and incorporate the sparsity assumption as an extra constraint for defining a worst case distribution.

Theorem 33. *For any $n, m, s \in \mathbb{N}^*$ such that n/m, n/s, $1 \vee (s/m)$ are integers, $n/m, n/s \geq 2$, any $\Delta \in (0, \frac{ms}{2(n-m)}]$ and any consistent policy, there is a problem with n arms — characterized by some action space $\mathcal{A}$ with $m = \max\{|A|,\ A \in \mathcal{A}\}$ and some vector of outcomes $\mathbf{X}$ with all gaps equal to Δ satisfying $\|\mathbf{X}\|_\infty \leq 1$, $\|\mathbf{X}\|_0 \leq s$ — on which the regret satisfies*

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

To give an idea of the proof, contrary to Theorem 31, we have more freedom in the covariance, and $\mathbf{X}$ can be chosen to maximize $\mathbb{V}(\sum_{i\in A} X_i)$ for each action A, up to the constraints $\|\mathbf{X}\|_\infty \leq 1$, $\|\mathbf{X}\|_0 = s$. The maximal value of $\sum_{i\in A} X_i$ is thus $(s \wedge m)$. Now consider for simplicity the softer constraint $\mathbb{E}\|\mathbf{X}\|_0 = s$. If $\mathbf{X}$ is chosen so that $\sum_{i\in A} X_i/(s \wedge m)$ is Bernoulli of parameter p, then the optimal p is equal to $(s \vee m)/n$. The variance is about $p(s \wedge m)^2 = ms(s \wedge m)/n$. Multiplying this by n/m (the number of actions) and dividing by the gap Δ gives the order of the lower bound.

Proof of Theorem 33. Consider $\mathcal{A}$ containing n/m disjoint actions $A_1, \dots, A_{n/m}$ composed of m arms. $\mathbf{X}$ is constructed as follows: $(1 \vee s/m)$ different actions are randomly chosen among $\mathcal{A}$, with equal probability, except the one for action A_1, that have an offset of δ. From

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

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

This problem reduces to a standard bandit problem with n/m Bernoulli arms. However, we have an additional piece of information, namely that the sum of the means is s. Thus, we can't apply the lower bound from Lai and Robbins (1985), since the distribution family has not a product form (changing the mean of one arm, we have to make sure that the sum of the means doesn't change, so we have to change at least another mean). Instead, we use the lower bound result from Graves and Lai

(1997) (Theorem 2), where we can increase the mean of one arm i while decreasing the mean of the others. Scaling the regret by $(s \wedge m)^{-1}$, we want to upper bound

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

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

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

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

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

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

Thus, we get the upper bounds

$$\mathrm{kl}\left(\left(\frac{m}{n}\vee\frac{s}{n}\right) - \delta\frac{m}{n}, \left(\frac{m}{n}\vee\frac{s}{n}\right) + \delta\left(1-\frac{m}{n}\right)\right) \leq \frac{4\delta^2}{\left(\frac{m}{n}\vee\frac{s}{n}\right)} \tag{7.4}$$

$$\mathrm{kl}\left(\left(\frac{m}{n}\vee\frac{s}{n}\right) - \delta\frac{m}{n} - \frac{\delta}{\frac{n}{m}-2}, \left(\frac{m}{n}\vee\frac{s}{n}\right) - \delta\frac{m}{n}\right) \leq \frac{4\delta^2}{\left(\frac{n}{m}-2\right)^2\left(\frac{m}{n}\vee\frac{s}{n}\right)}. \tag{7.5}$$

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

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

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

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

Using the bounds (7.4) and (7.5), we can relax the above constraint as

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

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

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

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

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

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

7.3.2 Our approach for sparse semi-bandits

In this subsection, we adapt our techniques to the sparse semi-bandit setting. Since $\|\mathbf{X}\|_\infty \leq 1$, the ℓ_0-inequality $\|\mathbf{X}\|_0 \leq s$ immediately implies the ℓ_1-inequality $\|\mathbf{X}\|_1 \leq s$. As we will actually only use sparsity through the latter inequality, we can relax our assumption on the model into $\|\mathbf{X}\|_1 \leq s$, for more generality. Let $\nu_i^* \triangleq \mathbb{E}[|X_i|]$, and $\overline{\nu}_{i,t-1}$ the corresponding empirical average estimate: $\overline{\nu}_{i,t-1} \triangleq \frac{\sum_{u\in[t-1]} \mathbb{I}\{i\in A_u\}|X_{i,u}|}{N_i t-1}$. Our approach is based on replacing $\sum_{j\in A} 0 \vee \Sigma_{ij}^*$ by $\nu_i^*(s \wedge m)$ (see Lemma 11). Using this, it is possible to estimate ν_i^* instead of each Σ_{ij}^*.

Lemma 11. $\sum_{j\in A} 0 \vee \Sigma_{ij}^* \leq 2\nu_i^*(s \wedge m)$.

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

$$\begin{aligned}
\sum_{j\in A} 0 \vee \Sigma_{ij}^* &= \sum_{j\in A} 0 \vee \mathbb{E}\left[X_i X_j - \mu_i^* \mu_j^*\right] \\
&\leq \sum_{j\in A}\left(\mathbb{E}[|X_i X_j|] + \left|\mu_i^* \mu_j^*\right|\right) \\
&= \mathbb{E}\left[|X_i| \sum_{j\in A}|X_j|\right] + |\mu_i^*| \sum_{j\in A}\left|\mu_j^*\right| \\
&\leq 2\mathbb{E}[|X_i|](s \wedge m).
\end{aligned}$$

□

We can therefore use the same algorithm (Algorithm 11), but with a confidence region $\mathcal{C}_t$ independent of A, since summing over A or $[n]$ on the main sum doesn't change the algorithm and the second sum $\sum_{j\in A} 0 \vee \Sigma_{ij,t}$ is replaced by an estimates of the upper bound given in Lemma 11.

$$\mathcal{C}_t \triangleq \overline{\boldsymbol{\mu}}_{t-1} + \left\{\boldsymbol{\xi} \in \mathbb{R}^n, \sum_{i\in[n]} \frac{N_i t - 1\xi_i^2}{m|\xi_i| + 2(s\wedge m)\nu_{i,t}} \leq 8(\log(t) + \log(\log(t))) + 4em\right\}, \quad (7.6)$$

where the upper bound estimate $\nu_{i,t} \triangleq \overline{\nu}_{i,t-1} + \sqrt{\frac{1.5\log(t)}{N_i t-1}}$ of ν_i^* is a simple consequence of Heoffding's inequality, using that $|X_{i,u}|$ is $1/4$-sub-Gaussian. Our algorithm is stated in Algorithm 12. As a byproduct of Theorem 32, we provide an upper bound for the regret in the sparse semi-bandit setting in Corollary 3 (see (7.7), for a more

Algorithm 12 ESCB-C modified for the case of $\|\cdot\|_1$-constrained outcomes

Initialization:
Play $A_1 = [n]$, or at least a sequence $A_1, A_2, \dots,$ (no more than n) such that all arm have been sampled once. We thus have $N_i t - 1 \geq 1$ for every arm $i \in [n]$.
For all subsequent rounds t:
Solve the following bilinear program to get A_t, with $\mathcal{C}_t$ define by (7.6), and play A_t.

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

precise bound). Again, notice we are reaching the lower bound of Theorem 33, using the relation $\sum_i \nu_i^* = \mathbb{E}\|\mathbf{X}\|_1 \leq s$.

Corollary 3. *Assume that the outcome distribution* $\mathbb{P}_\mathbf{X}$ *satisfies* $\|\mathbf{X}\|_\infty \leq 1$ *and* $\|\mathbf{X}\|_1 \leq s$, *and that*

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

for some universal constant c. *Then the regret of Algorithm 12 is upper bounded as*

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

where c' *is a universal constant.*

The corollary is obtained in the same way as Theorem 32. We can underline the difference that we don't have to construct n^2 covariance estimates, but only n (only the ν_i^*'s). In particular, as these estimates uses $1/4$-sub-Gaussian variables, we don't use the sub-exponential concentration of Lemma 10, which removes one term from the previous result. The obtained bound is

$$R_T \leq \Delta_{\max}\left(n + \frac{8nm^2}{\Delta^2} + c\right) + c' \log(m+1)\delta(T)\Bigg[\log(m+1) \sum_{i \in [n]} \frac{\nu_i^*(s \wedge m)}{\Delta_{i,\min}} + \sum_{i \in [n]} m\left(1 + \log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right) + \sum_{i \in [n]} \frac{(s \wedge m)^{2/3}}{\Delta_{i,\min}^{1/3}}\Bigg], \tag{7.7}$$

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

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

for some constant c, which gives our condition in Corollary 3.

Remark 16. *It should be noticed that semi-bandits algorithms as* CUCB-V *or* CUCB-KL *(that are variant of the classical* CUCB *(Kveton, Wen, Ashkan, and Szepesvari, 2015b), where the confidence region is a Cartesian product of confidence intervals, with Bernstein and kl-base confidence intervals respectively) also reach the lower bound of Theorem 33 for the regime $s \geq m$, since $\mathbb{V}(X_i) \leq 2\nu_i^*$ (thanks to Lemma 11). However, in the regime where $s \leq m$, these algorithms are not able to reach it, while* ESCB-C *is. In the following, we describe the two algorithms* CUCB-V *and* CUCB-KL*, and comment further on the tightness difference between confidence regions.*

7.3.3 Confidence regions comparison

We give here the two algorithms CUCB-V and CUCB-KL, which, as we have seen, also matches the lower bound given to the Theorem 33, in the specific regime where $s \geq m$. Both the two algorithms rely on the same optimization $A_t = \arg\max_{A \in \mathcal{A}} \mathbf{e}_A^\mathsf{T} \boldsymbol{\mu}_t$, where the vector $\boldsymbol{\mu}_t$ is defined for CUCB-V as

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

where

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

and for CUCB-KL as $\forall i \in [n], \mu_{i,t}$ is the unique solution x to

$$N_{i,t-1} \mathrm{kl}\Big(\overline{\mu}_{i,t-1}, x \Big) = \zeta \log(t) \text{ such that } x \in [\overline{\mu}_{i,t-1}, 1].$$

We take $\zeta = 1.2$ (although all $\zeta > 1$ are valid). The algorithms above can also be seen as a bilinear maximization where $\boldsymbol{\mu}_t$ is maximized over a confidence region that is a Cartesian product one 1-demendional confidence intervals. We illustrate in Figure 7.1 the difference between the confidence region considered in ESCB-C (when the correlation is low) and CUCB-KL. The red points represent $\boldsymbol{\mu}_t$ for each region. It can be seen that the Cartesian product confidence region greatly overestimates the risk in directions that are not close to the axes, giving rise to over-exploration. It is important to note however that this price to pay can be interesting in practice, because the corresponding algorithms are then very efficient (LP over $\mathcal{A}$, supposed possible[4]). As we noted in Remark 15, considering the intersection between the two confidence regions gives rise to an even tighter region, and therefore a better regret bound.

[4]Otherwise an approximation regret would be a more appropriate performance measure to consider.

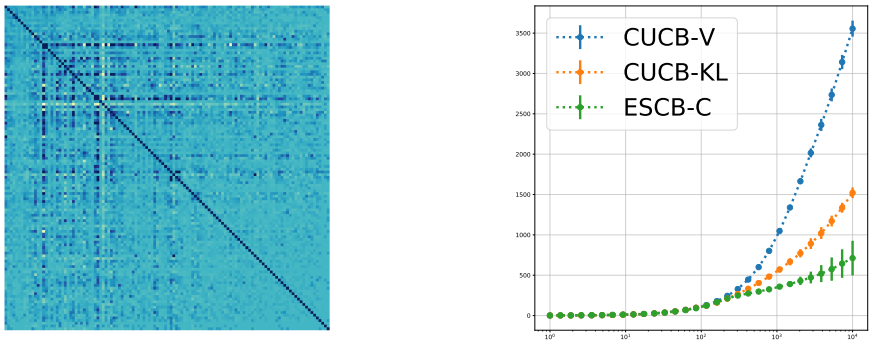

FIGURE 7.2: **Left**: Correlation matrix of the dataset, **right**: Cumulative regret, averaged over 36 independent simulations

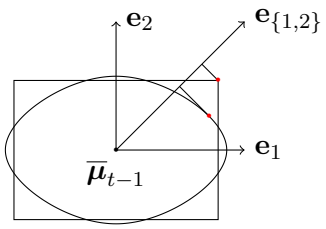

FIGURE 7.1: *Confidence regions built by* ESCB-C *(the pseudo-ellipse), and* CUCB-KL *(the rectangle), for* $\|\cdot\|_1$ *constrained outcomes. Notice that* CUCB-KL *has slightly better confidence intervals along the axis, but that* ESCB-C *is better in the direction* $\mathbf{e}_{\{1,2\}}$.

7.4 Experiments and discussion

We consider the following dynamic assortment problem. An agent has n products to sale, with fixed known prices. At each round, a customer arrives, with some unknown random valuation vector over products. Then, the agent offers any subset of products, by paying a fixed known cost for each offered product (e.g. transportation and display cost), and the customer buys an offered product if and only if its valuation is greater than its price. The agent is interested in maximizing the total profit (revenue minus cost) from sales over T rounds. We use the $n = 120$ products from the *Kaggle* (2013) dataset containing 7500 grocery store transactions. At each round, valuations are determined by sampling a random transaction from this dataset. The choice of such data is motivated by correlations that exist between arms, as illustrated in Figure 7.2 – left, representing the correlation matrix. We ran 36 independent simulations with $T = 10^4$, and with a common product price and cost respectively equal to 1.5 and 0.1. We compared CUCB-V and CUCB-KL with the Lovász extension implementation of ESCB-C (see subsection 7.5.1) and results are plotted in log-scale (Figure 7.2 – right); error bars represent the sample standard deviation over simulations. There is less volatility in the regret of CUCB-V and CUCB-KL; this is due to the fact that their confidence regions overestimate the risk, and the "bad" event where the regret deviates is almost negligible. Nevertheless, we clearly observe that ESCB-C outperforms the two other approaches in terms of the average regret. Finally, let us point out that we did not empirically compare to the OLS-UCB algorithm of Degenne and Perchet (2016b) since it is inefficient to implement (the combinatorial problem to be solved within each round is NP-Hard in general (Atamtürk and Gómez, 2017). We noticed

that for the choice of sub-Gaussianity matrix where all the correlation coefficients equals 1, OLS-UCB (if it could be implemented) would return a solution very close to CUCB-V.

7.4.1 Discussion

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

7.5 Appendix

7.5.1 Implementation using the Lovász extension

We now discuss the computational efficiency of our approaches. First, Algorithm 11 (and both those of Combes et al. (2015) and Degenne and Perchet (2016b)) is not efficient for arbitrary combinatorial space $\mathcal{A}$. However, the evaluation of $F : A \mapsto \max_{\boldsymbol{\mu}\in\mathcal{C}_t(A)} \mathbf{e}_A^\intercal \boldsymbol{\mu}$, can be done efficiently as it is an LP over a convex set. In practice, when $\mathcal{A}$ allows it, GREEDY[5] (Nemhauser, Wolsey, and Fisher, 1978) can be used to maximize F. In general, it is unknown if this alters the regret rate. On the one hand, it does not when $\mathcal{A}$ is given by a matroid, and $\mathcal{C}_t$ is as in Algorithm 12. This is because F is *submodular* and the following approximation guaranty holds for the output A_t of GREEDY (see Chapter 5): $2\Big(F(A_t) - \mathbf{e}_{A_t}^\intercal \overline{\boldsymbol{\mu}}_{t-1}\Big) + \mathbf{e}_{A_t}^\intercal \overline{\boldsymbol{\mu}}_{t-1} \geq F(A^*)$, where the l.h.s. is simply F where $\mathcal{C}_t$ is scaled by a factor 2 from its center $\overline{\boldsymbol{\mu}}_{t-1}$. On the other hand, when $\mathcal{C}_t(A)$ is as in Algorithm 11, a concave extension of $A \mapsto F(A)$ can be considered, and can thus be maximized efficiently. Notice, when considering the intersection of the two confidence regions as in Remark 15, this implementation is still tractable since the minimum of two concave functions is still concave. Since the obtained solution might not be fractional, we use a randomized rounding to obtain a feasible set $A_t \in \mathcal{A} = \{0,1\}^n$. We provide in the following further details and prove that this method scales the regret by a factor $1 + \log\Big(\frac{m\log(T)}{\Delta^2}\Big)$, an acceptable price for efficiency.

From the step 1 of the proof of Theorem 32, we have that

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

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

Since the final bound of Theorem 32 relies on the above upper bound, in Algorithm 11, instead of maximizing $A \mapsto \max_{\boldsymbol{\mu} \in \mathcal{C}_t(A)} \mathbf{e}_A^\intercal \boldsymbol{\mu}$, we can maximize

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

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

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

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

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

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

where G_1 and G_2 are the supermodular functions corresponding to the second and last term respectively. Then, using an ℓ_2 upper bound on G_i, $i \in \{1,2\}$ the probabilities p_j can be aggregated to build triggering probabilities. Since the triggering probabilities are inside the square root in the above bonus, we can apply Theorem 16. This gives a regret bound with an extra factor of $1 + \log\left(\frac{m \log(T)}{\Delta^2}\right)$. Notice, an application of Theorem 15 is also possible, but the extra factor would depend on $\boldsymbol{p}$ and would be difficult to bound.

Remark 17. *We addressed the efficiency by using randomization. Notice, this randomization is an example of "undergone randomization", since the selected action S is actually $\mathbf{x}$, that well describe the distribution used to generate the played set of arms. Saying it differently, notice that the initial setting was into the CMAB framework, whereas we used triggering probabilities to solve it. The fact that randomization has been transformed into triggering probabilities illustrates that the action of interest is not the random set played but the vector $\mathbf{x}$ which is the result of the continuous optimization of the objective considered above.*

In Chapter 8, we will see an example where the opposite happens: the randomization is controlled by the agent. More precisely, this policy will be based on Thompson sampling, which is an example where the chosen action is the one of interest, because it is the solution of a maximization problem where the parameters are random.

7.5.2 Proof of Theorem 32

Proof of Theorem 32. Let $t \geq 1$, and $\delta(t) \triangleq 2(\log(t) + \log(\log(t))) + em$. Through initialization, we can assume $N_{ij,t-1} \geq 1$ for all $i, j \in [n]$ (as this only adds $n(n-1)\Delta_{\max}/2$ to the regret bound). We will decompose contributions to regret by considering the following events:

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

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

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

We also define

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

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

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

and

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

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

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

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

i.e., using Cauchy-Schwarz and $\boldsymbol{\mu}_t \in C_t(A_t)$,

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

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

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

$$
\begin{aligned}
&\leq 4\sqrt{\delta(t)\sum_{i\in A_t}\frac{\sum_{j\in A_t} 0\vee\Sigma_{ij,t}}{N_{i,t-1}}} + \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \\
&\leq 4\sqrt{\delta(t)\sum_{i\in A_t}\frac{\sum_{j\in A_t} 0\vee\left(\Sigma^*_{ij}+2g_{ij}(t)\right)}{N_{i,t-1}}} + \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \qquad \mathfrak{S}_t \\
&\leq 4\sqrt{\delta(t)\sum_{i\in A_t}\frac{\sum_{j\in A_t} 0\vee\Sigma^*_{ij}}{N_{i,t-1}}} + 4\sqrt{\delta(t)\sum_{i\in A_t}\frac{\sum_{j\in A_t} 2g_{ij}(t)}{N_{i,t-1}}} \\
&+ \frac{8m\delta(t)}{\min_{j\in A_t} N_{j,t-1}} \\
&\leq \underbrace{4\sqrt{\delta(T)\sum_{i\in A_t}\frac{\max_{A\in\mathcal{A},i\in A}\sum_{j\in A} 0\vee\Sigma^*_{ij}}{N_{i,t-1}}}}_{(7.8)} + \underbrace{4\sqrt{\delta(T)\sum_{i,j\in A_t}\widetilde{g}_{ij}(T)}}_{(7.9)} \\
&+ \underbrace{\frac{8m\delta(T)}{\min_{j\in A_t} N_{j,t-1}}}_{(7.10)}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{I}\{\Delta(A_t)\leq (7.8)+(7.9)+(7.10)\} \\
&\leq \mathbb{I}\{\Delta(A_t)/3\leq (7.8)\}+\mathbb{I}\{\Delta(A_t)/3\leq (7.9)\}+\mathbb{I}\{\Delta(A_t)/3\leq (7.10)\}.
\end{aligned}
$$

We apply Theorem 13, that is helpful to bound the regret on each of this 3 events. Indeed, for the first term, applying it with $\beta_{i,T}=12^2\delta(T)\max_{A\in\mathcal{A},i\in A}\sum_{j\in A}0\vee\Sigma^*_{ij}$ and $\alpha_i=1/2$ gives the bound

$$
\sum_{t=1}^{T}\mathbb{I}\{\Delta(A_t)/3\leq (7.8)\}\Delta(A_t)\leq 4608\log_2^2(4\sqrt{m})\sum_{i\in[n]}\frac{\delta(T)\max_{A\in\mathcal{A},i\in A}\sum_{j\in A}0\vee\Sigma^*_{ij}}{\Delta_{i,\min}}.
$$

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

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

Thus, again, it is sufficient to treat each term separately. We also apply Theorem 13, but with $[n]^2$ as the set of arms, and with A_t^2 as the set of played arms, taking respectively $\alpha_i=1,\beta_{i,T}=24\sqrt{54+\sqrt{48}}\delta(T)$ and $\alpha_i=0.75,\beta_{i,T}=192\cdot 6^{2/3}\delta(T)$ for each term. This gives

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{I}\{\Delta(A_t)/3\leq (7.9)\}\Delta(A_t) \;&\leq 1152\sqrt{6}\log_2(4m)\sum_{i,j\in[n]}\delta(T)\left(1+\log\left(\frac{\Delta_{ij,\max}}{\Delta_{ij,\min}}\right)\right) \\
&+ 12288\cdot 6^{2/3}\left(4^{1/3}-1\right)^{-1}\log_2(4m)\sum_{i,j\in[n]}\delta(T)\Delta_{ij,\min}^{-1/3}.
\end{aligned}
$$

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

$$(7.10) \leq 8m\delta(T)\sqrt{\sum_{i\in A_t}\frac{1}{N_{i,t-1}^2}}.$$

Then, taking $\alpha_i = 1, \beta_{i,T} = 24m\delta(T)$ in Theorem 13 gives

$$\sum_{t=1}^{T}\mathbb{I}\{\Delta(A_t)/3 \leq (7.10)\}\Delta(A_t) \leq 1152\log_2(4\sqrt{m})\sum_{i\in[n]} m\delta(T)\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right).$$

This concludes step 1; notice that all subsequent steps will aim to bound the regret by a term independent of T, over a certain event. Thus, we can see that the bounds above are the actual contributions to the rate of the regret. To show Theorem 32, we must therefore choose the regime for $\Delta \leq \Delta_{i,\min}$ so that the first term prevails over the others. In other words, we want to have

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

where c is a constant. This gives exactly our condition in Theorem 32.

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

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

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

Notice that $\mathfrak{S}_t, \neg\mathfrak{C}_t$ implies $\mathfrak{A}_t$ for $\delta = \delta(t)$. Since each number of pulls $N_{i,t-1}$ for $i \in A^*$ is bounded by t, the number of possible $\mathbf{d} \in (\mathbb{N}^*)^{A^*}$ such that $\mathbb{P}[\mathfrak{B}_{\mathbf{d},t}] > 0$ is bounded by $(\log(t)/\log(\delta/(\delta-1)))^m$. Thanks to the following Lemma 12, and an union bound on those possible $\mathbf{d} \in (\mathbb{N}^*)^{A^*}$, we get

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

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

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

that is bounded by

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

Lemma 12 (Covering-argument)**.** *Let* $\mathbf{d} \in (\mathbb{N}^*)^{A^*}$*. Then,*

$$\mathbb{P}[\mathfrak{A}_t \cap \mathfrak{B}_{\mathbf{d},t}] \leq \left(\frac{(\delta-1)e}{m}\right)^m e^{1-\delta}.$$

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

$$\bigcap_{i\in A^*} \left\{ N_{i,t-1} \frac{\left(0 \vee \left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)^2}{4\left(\sigma_i^2 + |A^*|\left(\mu_i^* - \overline{\mu}_{i,t-1}\right)\right)} > \zeta_i \right\}$$

implies

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

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

$$\bigcap_{i\in A^*} \left\{ \overline{\mu}_{i,t-1} \leq x_i\left(\frac{\delta}{\delta-1}\right)^{d_i} \right\}. \tag{7.11}$$

With $\epsilon_i \triangleq \mu_i^* - x_i\left(\frac{\delta}{\delta-1}\right)^{d_i}$ and $\lambda_i \triangleq \frac{\epsilon_i}{2\left(\sigma_i^2+|A^*|\varepsilon_i\right)}$, $i \in A^*$, this further implies:

$$\begin{aligned}
&\left(\frac{\delta}{\delta-1}\right)^{-1} \sum_{i\in A^*} \zeta_i \\
&= \sum_{i\in A^*} \left(\frac{\delta}{\delta-1}\right)^{d_i-1} \frac{\epsilon_i^2}{4(\sigma_i^2 + |A^*|\varepsilon_i)} && x_i(e^{d_i}) > -\infty, \\
&\leq \sum_{i\in A^*} N_{i,t-1} \frac{\epsilon_i^2}{4(\sigma_i^2 + |A^*|\varepsilon_i)} && \mathfrak{B}_{\mathbf{d},t} \\
&= \sum_{i\in A^*} N_{i,t-1} \frac{\epsilon_i^2}{2(\sigma_i^2 + |A^*|\varepsilon_i)} - \sum_{i\in A^*} N_{i,t-1} \frac{\epsilon_i^2}{4(\sigma_i^2 + |A^*|\varepsilon_i)} \\
&\leq \sum_{i\in A^*} N_{i,t-1} \frac{\epsilon_i^2}{2(\sigma_i^2 + |A^*|\varepsilon_i)} - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2 \frac{\varepsilon_i^2}{4(\sigma_i^2 + |A^*|\varepsilon_i)^2} && \frac{\sigma_i^2}{\sigma_i^2 + |A^*|\varepsilon_i} \leq 1, \\
&= \sum_{i\in A^*} N_{i,t-1}\lambda_i\varepsilon_i - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2\lambda_i^2 \\
&\leq \sum_{i\in A^*} N_{i,t-1}\lambda_i\left(\mu_i^* - \overline{\mu}_{i,t-1}\right) - \sum_{i\in A^*} N_{i,t-1}\sigma_i^2\lambda_i^2 && \text{using (7.11)},
\end{aligned}$$

and this last quantity is equal to

$$\sum_{u\in[t-1]}\Big((\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})^\mathsf{T}(\boldsymbol{\mu}^*-\mathbf{X}_u)-(\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})^\mathsf{T}\mathbf{D}(\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*})\Big),$$

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

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

Thus, we have

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

where the last inequality uses Assumption 3 and $\|\boldsymbol{\lambda}\odot\mathbf{e}_{A_u\cap A^*}\|_1\leq 1/2$. Now, notice that

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

equals

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

so from Markov inequality, we get the following bound:

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

thus, we showed that

$$\mathbb{P}\left[\mathfrak{B}_{\mathbf{d},t}\cap\bigcap_{i\in A^*}\left\{N_{i,t-1}\frac{\Big(0\vee\Big(\mu_i^*-\overline{\mu}_{i,t-1}\Big)\Big)^2}{4\Big(\sigma_i^2+|A^*|\Big(\mu_i^*-\overline{\mu}_{i,t-1}\Big)\Big)}>\zeta_i\right\}\right]\leq e^{-\sum_{i\in A^*}\zeta_i\left(\frac{\delta}{\delta-1}\right)^{-1}},$$

i.e.

$$\mathbb{P}\left[\bigcap_{i\in A^*}\left\{\mathbb{I}\{\mathfrak{B}_{\mathbf{d},t}\}N_{i,t-1}\frac{\left(0\vee\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)^2}{4\left(\sigma_i^2+|A^*|\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)}>\zeta_i\right\}\right]\leq e^{-\sum_{i\in A^*}\zeta_i e^{-1}},$$

By Lemma 8 of Magureanu, Combes, and Proutiere (2014), since $\delta\geq em$, we have

$$\begin{aligned}\mathbb{P}[\mathfrak{B}_{\mathbf{d},t}\cap\mathfrak{A}_t]&=\mathbb{P}\left[\mathfrak{B}_{\mathbf{d},t}\cap\left\{\sum_{i\in A^*}N_{i,t-1}\frac{\left(0\vee\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)^2}{4\left(\sigma_i^2+|A^*|\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)}\geq\delta\right\}\right]\\&=\mathbb{P}\left[\sum_{i\in A^*}\mathbb{I}\{\mathfrak{B}_{\mathbf{d},t}\}N_{i,t-1}\frac{\left(0\vee\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)^2}{4\left(\sigma_i^2+|A^*|\left(\mu_i^*-\overline{\mu}_{i,t-1}\right)\right)}\geq\delta\right]\\&\leq\left(\frac{(\delta-1)e}{m}\right)^m e^{1-\delta}.\end{aligned}$$

□

Step 3: If $\neg\mathfrak{D}_t$ hold The regret under this event can be bounded by $8nm^2\Delta_{\max}/\Delta^2$ using Proposition 14.

Step 4: If $\neg\mathfrak{S}_t$ hold From Proposition 25, the regret under this event is bounded by a universal constant.

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

$$\begin{aligned}R_T\leq\Delta_{\max}&\left(\frac{n(n-1)}{2}+\frac{8nm^2}{\Delta^2}+c\right)\\&+c'\log(m+1)\delta(T)\left[\log(m+1)\sum_{i\in[n]}\frac{\max_{A\in\mathcal{A},i\in A}\sum_{j\in A}0\vee\Sigma_{ij}^*}{\Delta_{i,\min}}\right.\\&+\sum_{i,j\in[n]}\kappa\left(1+\log\left(\frac{\Delta_{ij,\max}}{\Delta_{ij,\min}}\right)\right)\\&+\sum_{i\in[n]}m\kappa\left(1+\log\left(\frac{\Delta_{i,\max}}{\Delta_{i,\min}}\right)\right)\\&\left.+\sum_{i,j\in[n]}\frac{\kappa^{4/3}}{\Delta_{ij,\min}^{1/3}}\right].\end{aligned}\tag{7.12}$$

□

Chapter 8

Statistical and Computational Efficiency of Thompson Sampling

In this chapter, which is based on our article Perrault, Boursier, et al. (2020), we study an interesting alternative to the optimistic methods considered so far, namely Thompson sampling (TS). Apart from its empirical superiority (Chapelle and Li, 2011), TS is interesting in the context of CMAB because the played action is easy to compute: there isn't an additive bonus to the main objective, contrary to the ℓ_2-based approaches. TS could thus answer the question of the existence of an efficient policy with an optimal asymptotic regret (up to a factor poly-logarithmic with the action size), which 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.

8.1 A trade-off between optimality and computational efficiency?

In this chapter, we assume that the reward, given the choice of A_t, is a function of $\boldsymbol{\mu}^* \odot \mathbf{e}_{A_t}$. As we already saw, 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[1] (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

[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.

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

TABLE 8.1: 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 34, Theorem 35, Theorem 36 related to CTS-BETA, CTS-GAUSSIAN, CLIP CTS-GAUSSIAN respectively.

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, Cesa-Bianchi, and Fischer, 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, Wang, and Yuan, 2013; Kveton, Wen, Ashkan, and Szepesvari, 2015b), 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 8.1 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 4), 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, Perchet, and Valko (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 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, Varaiya, and Walrand, 1987; Kveton, Wen, Ashkan, Eydgahi, et al., 2014; Talebi and Proutiere, 2016). Recently, Cuvelier, Combes, and Gourdin (2020) provided another approach for approximating ESCB for a wide variety of action spaces, including the matching bandit setting (Gai, Krishnamachari, and Jain, 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, Korda, and Munos (2012) and Agrawal and Goyal (2012b) 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 an attractive policy because its time complexity is $\mathcal{O}(\text{poly}(n))$ under Assumption 4. 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:

$$\mathcal{O}\left(\frac{n\sqrt{m}\log(T)}{\Delta}\right) \text{ in the linear case and } \mathcal{O}\left(\frac{nm\log(T)}{\Delta}\right) \text{ in the general case.} \tag{8.1}$$

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 here.

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, Mannor, and Mansour (2014), that provided a frequentist high-probability regret bounds for TS with a general action space and a general feedback model — Komiyama, Honda, and Nakagawa (2015), that investigated TS for the m-sets action space — Wen, Kveton, and Ashkan (2015), that studied TS for contextual CMAB problems, using the Bayesian regret metric (see also Russo and Van Roy (2016)).

8.1.1 Contributions

We first improve the result of Wang and Chen (2018) by providing the regret upper bound $\mathcal{O}\left(\log^2(m)n\log(T)/\Delta\right)$ 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-Gaussianity 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, 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 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 8.1.

8.1.2 Model

The CMAB problem we consider is formally introduced as follows. 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 $\mathcal{H}_t \triangleq \sigma(\mathbf{X}_1 \odot \mathbf{e}_{A_1}, \dots, \mathbf{X}_{t-1} \odot \mathbf{e}_{A_{t-1}})$ and a possible extra source of randomness (as usual, 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 4. *The agent has access to an oracle with a time complexity $\mathcal{O}(\text{poly}(n))$ such that for any mean vector $\boldsymbol{\mu}$,* $\text{Oracle}(\boldsymbol{\mu}) \in \arg\max_{A \in \mathcal{A}} r(A; \boldsymbol{\mu})$.

As in Chen, Wang, and Yuan (2016), we assume that the function r satisfies the following smoothness property.

Assumption 5. *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 minimum size of an optimal action $m^* \triangleq \min_{A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\intercal \boldsymbol{\mu}^*} |A|$.

8.2 The independent case: Thompson sampling with beta prior

In this section, we consider the following assumption on top of the CMAB setting from section 8.1.2.

Assumption 6. *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)).*

Algorithm 13 CTS-BETA

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

For this problem, we consider CTS-BETA in Algorithm 13, 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 $X_{i,t}$, it first binarizes it by sampling $Y_{i,t} \sim \mathrm{Bernoulli}(X_{i,t})$ (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 34, that improves the regret bound of Wang and Chen (2018) for CTS-BETA.

Theorem 34. *The policy π described in Algorithm 13 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 34, as well as the complete non-asymptotic upper-bound is postponed to subsection 8.5.1. 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 34 in subsection 8.5.1) 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 34 (compared to Equation 8.1) is a consequence of two ingredients. The first is the following concentration inequality (see Lemma 14), 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.

$$\mathbb{P}\left[\left\|\mathbf{e}_{A_t} \odot (\boldsymbol{\theta}_t - \overline{\boldsymbol{\mu}}_{t-1})\right\|_1 \geq \sqrt{\frac{1}{2}\log(|\mathcal{A}|2^m T)\sum_{i\in A_t}\frac{1}{N_{i,t-1}}}\,\middle|\,\mathcal{H}_t\right] \leq 1/T. \tag{8.2}$$

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 34 in subsection 8.5.1) 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 $\mathcal{O}\left(\varepsilon^{-4m^*-2}\right)$, 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.

8.3 The general case: Thompson sampling with Gaussian prior

Algorithm 14 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 $\boldsymbol{\theta}_t \sim \otimes_{i\in[n]}\mathcal{N}\left(\overline{\mu}_{i,t-1}, N_{i,t-1}^{-1}\beta D_i\right)$ ($\theta_{i,t} \sim \mathcal{U}[a,b]$ 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.

In this section, we consider the setting from section 8.1.2, 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 26 gives two examples included in this assumption.

Assumption 7. *There exists a vector* $\mathbf{D} \triangleq (D_1,\dots,D_n) \in \mathbb{R}_+^n$ *known to the agent such that*

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

Proposition 26. *Assumption 7 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-Gaussianity matrix* $\mathbf{C}$ *(setting (iii)), taking* $D_i = \max_{A\in\mathcal{A},\ i\in A}\sum_{j\in A}|C_{ij}|$.

Proof. Assumption 7 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 $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right]$ is bounded by

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

The case of $\mathbf{C}$-sub-Gaussian outcomes with a known sub-Gaussianity matrix $\mathbf{C}$ (i.e., $\mathbb{E}\left[e^{\boldsymbol{\lambda}^\intercal(\mathbf{X}-\boldsymbol{\mu}^*)}\right] \le e^{\boldsymbol{\lambda}^\intercal\mathbf{C}\boldsymbol{\lambda}/2}$ for all $\boldsymbol{\lambda}\in\mathbb{R}^n$) is also captured, taking[2]

$$D_i = \max_{A\in\mathcal{A},\ i\in A}\sum_{j\in A}|C_{ij}|.$$

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

Indeed, for an action A,

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

□

For the above setting, we provide CTS-GAUSSIAN in Algorithm 14, 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 13 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 35.

Theorem 35. *The policy π described in Algorithm 14 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 35, as well as the complete non-asymptotic upper-bound is postponed to subsection 8.5.4. Nonetheless, in the following paragraphs, we provide some insights and highlight the novelty of our analysis. Notice, $\beta > 1$ is an artefact of the analysis and can in practice be taken equal to 1 (as we will do in our experiments).

Main proof challenges In the setting of the previous section, the outcomes are independent in $[0,1]$ and an important step in Algorithm 13 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 13. 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 happens. We let $\mathfrak{T}_t(Z)$ denote the complementary event. As for the proof of Theorem 34, we can condition on the history to rewrite this expected number of rounds and then upper bound it as

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

[3] In practice, for **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.

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 $\mathbb{E}\Big[\sup_{t\geq 1} \prod_{i\in Z'} g_i\Big(\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|\Big)\Big]$, where g_i are non-negative increasing functions. Our approach is to prove that $\Big(\left|\overline{\mu}_{i,t-1} - \mu_i^*\right|\Big)_i$ is *weakly stochastically dominated* by $\Big(\sqrt{\frac{\beta D_i}{N_{i,t-1}}}|\eta_i|\Big)_i$, where $\boldsymbol{\eta} \sim \otimes_i \mathcal{N}(0,1)$, which is the same vector but where the empirical mean is built with independent Gaussian outcomes instead. Notice, independence is crucial to be able to factorize the expectation $\mathbb{E}[\prod_{i\in Z'} g_i]$, in the same way as in the proof of Theorem 34. 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, $\mathbb{E}[\prod_i f_i(U_i)] \leq \mathbb{E}[\prod_i f_i(V_i)]$.
- 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 7 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 14, 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 34 (recall that Algorithm 13 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 (8.3) in Step 4 of the proof of Theorem 34 in subsection 8.5.1 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.

8.3.1 clip cts-gaussian for the linear reward case

In this subsection, we make the following assumptions on top of section 8.1.2.

Assumption 8. *The reward function is linear, defined as* $r(A, \boldsymbol{\mu}) \triangleq \mathbf{e}_A^\intercal \boldsymbol{\mu}$.

Assumption 9. *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 9 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 36. *The policy CLIP CTS-GAUSSIAN has regret of order*

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

Not only D_i is improved through the above relaxation, but also, the leading term is never worst than the one of CUCB. The proof and the complete non-asymptotic upper-bound is delayed to subsection 36. 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 $\text{Oracle}(\boldsymbol{\mu}) = \text{Oracle}\left(\boldsymbol{\mu} + \boldsymbol{\delta} \odot \mathbf{e}_{\text{Oracle}(\boldsymbol{\mu})}\right)$, 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 36 is comparable to the one for OLS-UCB from Degenne and Perchet (2016b). Indeed, we recall that they obtained a factor of order $\Gamma_{ii}\left((1-\gamma)\log^2(m) + \gamma m\right)$, with γ as defined in the introduction of the thesis (i.e., $\gamma = \max_{A\in\mathcal{A}} \max_{(i,j)\in A^2, i\neq j} (0 \vee \Gamma_{ij})/\sqrt{\Gamma_{ii}\Gamma_{jj}}$), where we have $\left(D_i \log^2(m) \wedge m\Gamma_{ii}\right)$. 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.

8.4 Experiments and discussion

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 8.1. 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, Brändén, and Liggett (2009), corollary 4.18). It is clear from the experiments that CTS policies

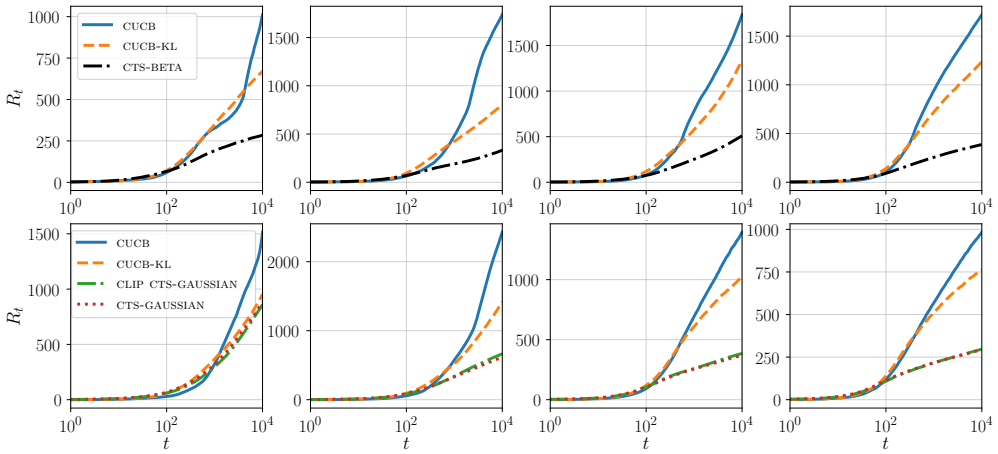

FIGURE 8.1: Cumulative regret (averaged over 50 simulations) for the shortest path problem. **Top:** with mutually independent outcomes, taking the opposite sum of means being $s = 70, 90, 110, 130$ respectively. **Bottom:** with correlated outcomes, taking the opposite sum of outcomes being $s = 70, 90, 110, 130$ respectively.

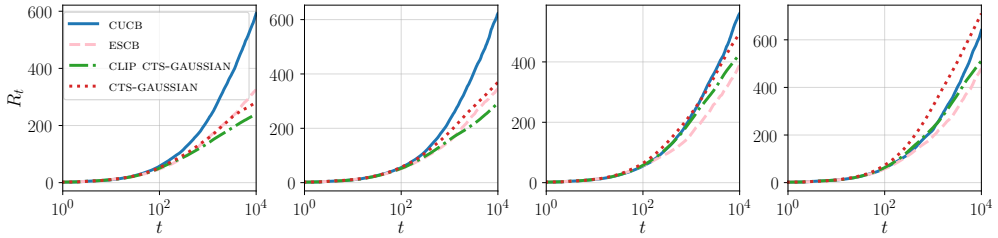

FIGURE 8.2: Cumulative regret (averaged over 50 simulations) for the matching problem with Gaussian outcomes, taking $c = -1/n, 0.2, 0.5, 1$ respectively.

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 8.2, 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 8.2), and observe the efficiency of CUCB and CLIP CTS-GAUSSIAN compared to ESCB.

	$K_{3,3}$	$K_{4,4}$	$K_{5,5}$	$K_{6,6}$	$K_{7,7}$	$K_{8,8}$
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

TABLE 8.2: Computation time per round (ms), with $c = 0.3$, $T = 100$, averaged over 5 simulations.

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 8.3 where we also compared with a common prior policy approach (Agrawal, Avadhanula, 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 $\mathcal{A} = \{\{km+1, \dots, (k+1)m\} \mid k \in \{0, \dots, \frac{n}{m} - 1\}\}$ 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.

8.4.1 Conclusion and future work

In this work, 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 for use in the prior distribution, as we did in the previous chapter. 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 the sub-Gaussianity. This latter result would thus show that CTS-BETA dominates CUCB-KL, which is empirically observed.

8.5 Missing proofs

8.5.1 Proof of Theorem 34

Proof of Theorem 34. We first restate the complete non-asymptotic upper-bound as follows.

[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.

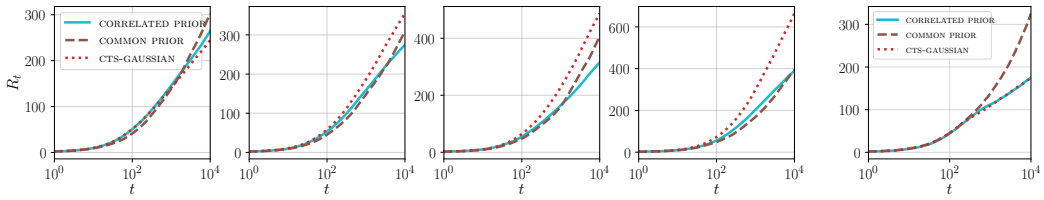

FIGURE 8.3: Comparison with correlated prior sampling and common prior sampling (averaged over 50 simulations). **The first 4:** for the $K_{4,4}$ matching problem, with Gaussian outcomes, taking $c = 0, 0.2, 0.5, 1$. **The last:** for $\mathcal{A} = \big\{\{km+1,\dots,(k+1)m\} \mid k \in \{0,\dots,\frac{n}{m}-1\}\big\}$, $c = 0$.

Theorem. *The policy π described in Algorithm 13 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 $\Delta_{\min}/(2B) - (m^{*2}+1)\varepsilon > 0$.

8.5.2 Preliminary lemmas

In order to prove Theorem 34, 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 13 (that is proved in the same way as Proposition 14).

Lemma 13. *In Algorithm 13, 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 14. *In Algorithm 13, for all round t, we have*

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

Proof. From (Marchal, Arbel, 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

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

□

8.5.3 Main proof

With the two lemmas from the previous subsection, we are ready to demonstrate Theorem 34. We consider the following events.

- $\mathfrak{Z}_t\triangleq\{\Delta_t>0\}$
- $\mathfrak{B}_t\triangleq\Big\{\exists i\in A_t,\ |A_t|\cdot\Big|\overline{\mu}_{i,t-1}-\mu_i^*\Big|>\Delta_{\min}/(2B)-(m^{*2}+1)\varepsilon\Big\}$
- $\mathfrak{C}_t\triangleq\Big\{\big\|\mathbf{e}_{A_t}\odot(\boldsymbol{\theta}_t-\boldsymbol{\mu}^*)\big\|_1>\Delta_t/B-\Big(m^{*2}+1\Big)\varepsilon\Big\}$
- $\mathfrak{D}_t\triangleq\Big\{\big\|\mathbf{e}_{A_t}\odot(\boldsymbol{\theta}_t-\overline{\boldsymbol{\mu}}_{t-1})\big\|_1\geq\sqrt{0.5\cdot\log(|\mathcal{A}|2^mT)\sum_{i\in A_t}1/N_{i,t-1}}\Big\}$.

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 13,

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

Step 2: bound under $\mathfrak{Z}_t\wedge\neg\mathfrak{B}_t\wedge\mathfrak{C}_t\wedge\mathfrak{D}_t$ By Lemma 14,

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

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

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

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 13 to get the bound

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

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

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

$$\mathfrak{S}_t(Z) \triangleq \{\forall \boldsymbol{\theta}' \text{ s.t. } \|(\boldsymbol{\mu}^* - \boldsymbol{\theta}') \odot \mathbf{e}_Z\|_\infty \leq \varepsilon,\ \mathfrak{R}(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \boldsymbol{\theta}_t \odot \mathbf{e}_{Z^c}, Z) \text{ holds}\} \quad (8.3)$$

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

We can state the three following lemmas. Note that Lemma 15 is exactly the Lemma 1 from Wang and Chen (2018). The other two replace their Lemma 7.

Lemma 15. *In Algorithm 13, 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 16. *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 13, 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 17. *In Algorithm 13, we have*

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

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 15 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 16 and 17 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 16. 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 τ_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 17. 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

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

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 $\tau \ge \tau_q + 1$, $N_{i,\tau-1} \ge q$ 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

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

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

$$\begin{aligned}
&\le \sum_{Z' \subset Z,\ Z' \ne \emptyset} \prod_{i \in Z'} \sum_{k_i \ge q} \left(6c'' \cdot e^{-\varepsilon^2 k/2}\varepsilon^{-2} + 2e^{-\varepsilon^2 k/8}\left(1 - 2e^{-\varepsilon^2 k/8}\right)^{-1}\right) \\
&\le e^{-\varepsilon^2 q/8}\left(c'\varepsilon^{-4}\right)^{|Z|},
\end{aligned}$$

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

$$\begin{aligned}
&\le \sum_{Z' \subset Z,\ Z' \ne \emptyset} \prod_{i \in Z'} \left(\sum_{k=q}^{\lfloor 8/\varepsilon^2 \rfloor} \left(4/\varepsilon^2 - 1\right) + \sum_{k \ge \lfloor 8/\varepsilon^2 \rfloor + 1}^{\infty} \left(6c \cdot e^{-\varepsilon^2 k/2}\varepsilon^{-2} + \frac{2e^{-\varepsilon^2 k/8}}{1 - 2e^{-\varepsilon^2 k/8}}\right)\right) \\
&\le \left(c\varepsilon^{-4}\right)^{|Z|},
\end{aligned}$$

where c, c' are two universal constant. □

□

8.5.4 Proof of Theorem 35

Proof of Theorem 35. We beginning by stating the complete version of Theorem 35.

Theorem. *The policy π described in Algorithm 14 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 35, we consider the same events as in the proof of Theorem 34, except for the event $\mathfrak{D}_t$, that becomes

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

Step 1 is unchanged. Step 2 and Step 3 are modified only through the event $\mathfrak{D}_t$, using the following modification of Lemma 14.

Lemma 18. *In Algorithm 14, 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

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

we have

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

□

The final bound on the regret in Step 3 is obtained using the same derivation as in Theorem 34, 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 $\mathbb{P}[\,|\theta_{i,t} - \mu_i^*| \le \varepsilon | \mathcal{H}_t]$ 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 34, to obtain the following upper bound, using $\mathfrak{M}_t$ to be able to start from $q = 1$.

$$\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}[\Delta(A_t)\mathbb{I}\{\mathfrak{M}_t \wedge \mathfrak{Z}_t \wedge \neg\mathfrak{C}_t\}] \\
&\le \sum_{Z\subset A^*,\ Z\neq\emptyset} \sum_{q\ge 1} \mathbb{E}\left[\sup_{\tau\ge\tau_q+1} \sum_{Z'\subset Z,\ Z'\neq\emptyset} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \le \varepsilon|\mathcal{H}_\tau]} - 1\right)\right] \\
&\le \sum_{Z\subset A^*,\ Z\neq\emptyset} \sum_{q\ge 1} \underbrace{\sum_{Z'\subset Z,\ Z'\neq\emptyset} \mathbb{E}\left[\sup_{\tau\ge\tau_q+1} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \le \varepsilon|\mathcal{H}_\tau]} - 1\right)\right]}_{(8.4)}.
\end{aligned}$$

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:

$$(8.4) \le \sum_{Z'\subset Z,\ Z'\neq\emptyset} \sum_{\mathbf{k}\in[q..\infty)^{Z'}} (8.5),$$

where

$$(8.5) \triangleq \mathbb{E}\left[\sup_{\tau\ge\tau_q+1} \mathbb{I}\{\forall i \in Z',\ N_{i,\tau-1} = k_i\} \prod_{i\in Z'} \left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \le \varepsilon|\mathcal{H}_\tau]} - 1\right)\right].$$

One can notice that for all $i \in Z'$, all $k_i \ge q$, $\mathbb{I}\{N_{i,\tau-1} = k_i\}\left(\frac{1}{\mathbb{P}[\,|\theta_{i,\tau} - \mu_i^*| \le \varepsilon|\mathcal{H}_\tau]} - 1\right)$ is of the form $\mathbb{I}\{N_{i,\tau-1} = k_i\} g_i\left(\left|\overline{\mu}_{i,\tau-1} - \mu_i^*\right|\right)$, 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 $\mathcal{N}\left(0, \beta D_i N_{i,\tau-1}^{-1}\right)$, which is symmetric, so we have

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

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 $\left|\overline{\mu}_{i,\tau-1} - \mu_i^*\right|$ 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 $\mathbf{k} \in [q..\infty)^{Z'}$,

$$\mathbb{E}\left[\sup_{\tau\ge\tau_q+1} \prod_{i\in Z'} \left(\mathbb{I}\{N_{i,\tau-1} = k_i\} g_i\left(\left|\overline{\mu}_{i,\tau-1} - \mu_i^*\right|\right)\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sup_{\tau\geq\tau_q+1}\prod_{i\in Z'}\left(\mathbb{I}\{N_{i,\tau-1}=k_i\}\int_0^\infty \mathbb{I}\Big\{g_i\Big(\big|\overline{\mu}_{i,\tau-1}-\mu_i^*\big|\Big)\geq u_i\Big\}\mathrm{d}u_i\right)\right]\\
&\leq \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\mathbb{E}\left[\sup_{\tau\geq\tau_q+1}\prod_{i\in Z'}\mathbb{I}\{N_{i,\tau-1}=k_i\}\mathbb{I}\Big\{g_i\Big(\big|\overline{\mu}_{i,\tau-1}-\mu_i^*\big|\Big)\geq u_i\Big\}\right]\mathrm{d}\mathbf{u}\\
&= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\mathbb{E}\left[\prod_{i\in Z'}\mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\Big\{g_i\Big(\big|\overline{\mu}_{i,\tau^*-1}-\mu_i^*\big|\Big)\geq u_i\Big\}\right]\mathrm{d}\mathbf{u}, \qquad (8.6)
\end{aligned}
$$

where τ^* is the first time τ such that the event

$$
\mathbb{I}\Big\{\forall i\in Z',\ N_{i,\tau-1}=k_i \text{ and } g_i\Big(\big|\overline{\mu}_{i,\tau-1}-\mu_i^*\big|\Big)\geq u_i\Big\}
$$

holds, and is ∞ if it never holds.

$$
\begin{aligned}
(8.6) &= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\mathbb{E}\left[\prod_{i\in Z'}\mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\Big\{g_i\Big(\big|\overline{\mu}_{i,\tau^*-1}-\mu_i^*\big|\Big)\geq u_i\vee g_i(0)\Big\}\right]\mathrm{d}\mathbf{u}\\
&= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\mathbb{E}\left[\prod_{i\in Z'}\mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\Big\{\big|\overline{\mu}_{i,\tau^*-1}-\mu_i^*\big|\geq g_i^{-1}(u_i\vee g_i(0))\Big\}\right]\mathrm{d}\mathbf{u}\\
&= \int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\sum_{\mathbf{s}\in\{-1,1\}^{Z'}}(8.7)\mathrm{d}\mathbf{u},
\end{aligned}
$$

where $(8.7)\triangleq\mathbb{E}\Big[\prod_{i\in Z'}\mathbb{I}\{N_{i,\tau^*-1}=k_i\}\mathbb{I}\Big\{s_i\Big(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\Big)\geq g_i^{-1}(u_i\vee g_i(0))\Big\}\Big]$. If we define

$$
(8.8)\triangleq\frac{\exp\left(\sum_{i\in Z'}N_{i,\tau^*-1}\left(\frac{s_ig_i^{-1}(u_i\vee g_i(0))}{D_i}\Big(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\Big)-\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2}{2D_i}\right)\right)}{\exp\left(\sum_{i\in Z'}\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2k_i}{2D_i}\right)},
$$

then (8.7) is bounded by

$$
\begin{aligned}
&\mathbb{P}\Big[(8.8)\geq 1,(N_{i,\tau^*-1})_{i\in Z'}=\mathbf{k}\Big]\\
&\leq\mathbb{P}\left[\frac{\exp\left(\sum_{i\in Z'}N_{i,\tau^*-1}\left(\frac{s_ig_i^{-1}(u_i\vee g_i(0))}{D_i}\Big(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\Big)-\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2}{2D_i}\right)\right)}{\exp\left(\sum_{i\in Z'}\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2k_i}{2D_i}\right)}\geq 1\right]\\
&\leq\frac{\mathbb{E}\left[\exp\left(\sum_{i\in Z'}N_{i,\tau^*-1}\left(\frac{s_ig_i^{-1}(u_i\vee g_i(0))}{D_i}\Big(\overline{\mu}_{i,\tau^*-1}-\mu_i^*\Big)-\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2}{2D_i}\right)\right)\right]}{\exp\left(\sum_{i\in Z'}\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2k_i}{2D_i}\right)}\\
&=\frac{\mathbb{E}\left[\exp\left(\sum_{t=1}^{\tau^*-1}\sum_{i\in Z'\cap A_t}\left(\frac{s_ig_i^{-1}(u_i\vee g_i(0))}{D_i}(X_{i,t}-\mu_i^*)-\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2}{2D_i}\right)\right)\right]}{\exp\left(\sum_{i\in Z'}\frac{\big(g_i^{-1}(u_i\vee g_i(0))\big)^2k_i}{2D_i}\right)}.
\end{aligned}
$$

From Assumption 7, 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, because $\mathbb{E}[M_\tau|\mathcal{F}_{\tau-1}]/M_{\tau-1}$ is equal to

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

Since τ^* is a stopping time with respect to $\mathcal{F}_\tau$, we have from Doob's optional sampling theorem for non-negative supermartingales[5] that $\mathbb{E}[M_{\tau^*}]\leq 1$. Therefore,

$$(8.7) \leq \exp\left(-\sum_{i\in Z'}\frac{\left(g_i^{-1}(u_i\vee g_i(0))\right)^2 k_i}{2D_i}\right).$$

Now, we want to use the following fact (see Chang, Cosman, and Milstein (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,

$$\begin{aligned}
(8.6) &\leq \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\prod_{i\in Z'}\mathbb{P}\left[\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\geq g_i^{-1}(u_i\vee g_i(0))\right]\mathrm{d}\mathbf{u}\\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\prod_{i\in Z'}\mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\geq u_i\vee g_i(0)\right]\mathrm{d}\mathbf{u}\\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\int_{\mathbf{u}\in\mathbb{R}_+^{Z'}}\prod_{i\in Z'}\mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\geq u_i\right]\mathrm{d}\mathbf{u}\\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\prod_{i\in Z'}\int_0^\infty\mathbb{P}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\geq u_i\right]\mathrm{d}u_i\\
&= \left(\sqrt{\frac{\pi}{2e}}\frac{2\beta}{\sqrt{\beta-1}}\right)^{|Z'|}\prod_{i\in Z'}\mathbb{E}\left[g_i\left(\sqrt{\frac{\beta D_i}{k_i}}|\eta_i|\right)\right].
\end{aligned}$$

[5] 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^*.

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

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

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:

$$\begin{aligned}A_1 &\leq 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{1}{\int_{-\varepsilon_i}^{0} e^{-(1-\frac{1}{2\alpha})ux}\mathrm{d}u}\mathrm{d}x \\ &= 2\int_{\alpha\varepsilon_i}^{+\infty} \frac{(1-\frac{1}{2\alpha})x}{e^{(1-\frac{1}{2\alpha})\varepsilon_i x} - 1}\mathrm{d}x. \qquad (8.9)\end{aligned}$$

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

$$\begin{aligned}(8.9) &\leq \frac{2e^{\left(\alpha-\frac{1}{2}\right)\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1}\int_{\alpha\varepsilon_i}^{+\infty}\left(1-\frac{1}{2\alpha}\right)xe^{-(1-\frac{1}{2\alpha})\varepsilon_i x}\mathrm{d}x \\ &= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1}\frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2}\int_{(\alpha-\frac{1}{2})\varepsilon_i^2}^{+\infty} xe^{-x}\mathrm{d}x \\ &= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1}\frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2}\left[-(x+1)e^{-x}\right]_{(\alpha-\frac{1}{2})\varepsilon_i^2}^{\infty} \\ &= \frac{2e^{(\alpha-\frac{1}{2})\varepsilon_i^2}}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1}\frac{1}{(1-\frac{1}{2\alpha})\varepsilon_i^2}\left(\left(\alpha-\frac{1}{2}\right)\varepsilon_i^2 + 1\right)e^{-(\alpha-\frac{1}{2})\varepsilon_i^2} \\ &= \frac{2}{e^{(\alpha-\frac{1}{2})\varepsilon_i^2} - 1}\left(\alpha + \frac{\alpha}{(\alpha-\frac{1}{2})\varepsilon_i^2}\right) \\ &\leq 4e^{-\varepsilon_i^2/12}.\end{aligned}$$

Otherwise, we have

$$\begin{aligned}(8.9) &= \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2}\int_{\alpha\varepsilon_i^2}^{\infty}\frac{u}{e^{\left(1-\frac{1}{2\alpha}\right)u} - 1}\mathrm{d}u \\ &\leq \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2}\int_{0}^{\infty}\frac{u}{e^{\left(1-\frac{1}{2\alpha}\right)u} - 1}\mathrm{d}u\end{aligned}$$

$$= \frac{2(1-\frac{1}{2\alpha})}{\varepsilon_i^2} \frac{\pi^2}{6\left(1-\frac{1}{2\alpha}\right)^2}$$
$$\leq \frac{24\beta D_i}{\varepsilon^2}.$$

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 (8.6) is bounded by

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

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

$$\sum_{Z\subset A^*,\ Z\neq\emptyset}\sum_{q\geq 1}\sum_{Z'\subset Z,\ Z'\neq\emptyset}\sum_{\mathbf{k}\in[q..\infty)^{Z'}}(8.6) \leq \left(C\varepsilon^{-2}\beta\max_i D_i\right)\left(\frac{C'\beta}{\sqrt{\beta-1}}\varepsilon^{-4}\beta^2\max_i D_i^2\right)^{m^*}$$

so,

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

□

8.5.5 Proof of Theorem 36

Proof of Theorem 36. The regret bound of CLIP CTS-GAUSSIAN 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.

8.5.6 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

$$S \triangleq \Big\{ i \in [n],\ \Gamma_{ii} m(\log(T) + 4\log\log(T)) \geq 4\log_2^2(4\sqrt{m})\beta D_i \log(|A|2^m T) \Big\}.$$

We have the following lemma.

Lemma 19.

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

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

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

□

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

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

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 6, 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 12 and Theorem 13 to get the bound

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

8.5.7 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.

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

where $\mathfrak{R}(\boldsymbol{\theta}', Z)$ is the event

$$\left\{ \forall A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}(\boldsymbol{\theta}') : Z \subset A,\ \mathbf{e}_{\text{Oracle}(\boldsymbol{\theta}')}^\mathsf{T} \boldsymbol{\theta}' > \mathbf{e}_{A^*}^\mathsf{T} \boldsymbol{\mu}^* - (m^*(m^*+1)/2 + 1)\varepsilon \right\}.$$

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 20.

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

This allows us to consider the success probability $\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t]$ in the analysis. Notice however that $Z \subset \text{Oracle}\Big(\Big(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\Big) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\Big)$, that is guaranteed when $\mathfrak{S}_t(Z) \wedge \neg\mathfrak{T}_t(Z)$ holds, does not necessarily implies that $Z \subset \text{Oracle}\Big(\widetilde{\boldsymbol{\theta}}_t\Big)$. However, it turns out that the following (true) predicate

$$Z \subset A,\ \forall A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}}\Big(\Big(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\Big) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\Big)$$

implies

$$Z \subset A,\ \forall A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}}\Big(\widetilde{\boldsymbol{\theta}}_t\Big).$$

This fact is from Lemma 21, with $\boldsymbol{\eta} = \Big(\boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\Big) \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}$ and $\boldsymbol{\delta} = \Big(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\mu}^* \wedge \widetilde{\boldsymbol{\theta}}_t\Big) \odot \mathbf{e}_Z$.

Lemma 21. *Let $\boldsymbol{\eta} \in \mathbb{R}^n$, $\boldsymbol{\delta} \in \mathbb{R}^n_+$ such that for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}}\boldsymbol{\eta}$, we have $Z \subset A$. Then, for all $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^{\mathsf{T}}(\boldsymbol{\eta} + \boldsymbol{\delta} \odot \mathbf{e}_Z)$, 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

$$\mathbb{P}[\neg\mathfrak{T}_t(Z)|\mathcal{H}_t] = \mathbb{P}\Big[\forall i \in Z,\ \varepsilon \vee \Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big) - 0 \vee \Big(\theta_{i,t} - \overline{\mu}_{i,t-1}\Big) \leq \varepsilon \Big| \mathcal{H}_t\Big],$$

  is a product of functions that are decreasing with respect to $\varepsilon \vee \Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big)$.

- We use that $\varepsilon \vee \Big(\mu_i^* - \overline{\mu}_{i,t-1}\Big) \geq g_i^{-1}(u_i \vee g_i(\varepsilon))$ 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 7 with $\boldsymbol{\lambda} \in \mathbb{R}^n_+$.

Proof of Lemma 20. It is sufficient to prove that

$$\mathfrak{Z}_t, \neg\mathfrak{C}_t \Rightarrow \exists Z \subset A^*,\ Z \neq \emptyset \text{ s.t. } \mathfrak{S}_t(Z) \text{ holds,} \tag{8.10}$$

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

$$\begin{aligned}
\mathbf{e}_{A_t}^{\mathsf{T}}\widetilde{\boldsymbol{\theta}}_t &= \max_{A \in \mathcal{A}} \mathbf{e}_A^{\mathsf{T}}\widetilde{\boldsymbol{\theta}}_t \\
&\geq \max_{A \in \mathcal{A}} \mathbf{e}_A^{\mathsf{T}}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\Big) \\
&= \mathbf{e}_{\text{Oracle}\left(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\right)}^{\mathsf{T}}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_Z + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z^c}\Big) \\
&> \mathbf{e}_{A^*}^{\mathsf{T}}\boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon,
\end{aligned}$$

giving $\mathfrak{C}_t$. To prove (8.10), we first consider the choice $Z = Z_1 = A^*$. Two cases can be distinguished:

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

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

1a) For the first case, consider any vector $\boldsymbol{\theta}'$ such that $0 \leq (\boldsymbol{\mu}^* - \boldsymbol{\theta}') \odot \mathbf{e}_{A^*} \overset{(8.11)}{\leq} \varepsilon \mathbf{e}_{A^*}$ and let $A \overset{(8.12)}{=} \text{Oracle}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big)$. We can write

$$\mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) \overset{(8.13)}{\geq} \mathbf{e}_{A^*}^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) \overset{(8.14)}{\geq} \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - m^*\varepsilon,$$

where (8.13) is from (8.12), and (8.14) is from (8.11). This rewrites as

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

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 $0 \overset{(8.15)}{\leq} (\boldsymbol{\mu}^* - \boldsymbol{\theta}') \odot \mathbf{e}_{A^*} \overset{(8.16)}{\leq} \varepsilon \mathbf{e}_{A^*}$, and some action $A \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big)$ 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 $S' \cap A^* \overset{(8.17)}{=} \emptyset$, then $A \neq S'$:

$$\begin{aligned}
\mathbf{e}_{S'}^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) \overset{(8.18)}{=} \mathbf{e}_{S'}^\mathsf{T}\widetilde{\boldsymbol{\theta}}_t &\overset{(8.19)}{\leq} \mathbf{e}_{A_t}^\mathsf{T}\widetilde{\boldsymbol{\theta}}_t \\
&\overset{(8.20)}{\leq} \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - (m^*(m^*+1)/2+1)\varepsilon \\
&< \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - m^*\varepsilon \\
&\overset{(8.21)}{\leq} \mathbf{e}_{A^*}^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big),
\end{aligned}$$

where (8.18) is from (8.17), (8.19) is from the definition of A_t, (8.20) is from $\neg\mathfrak{C}_t$ and (8.21) is from (8.16). Now, we again distinguish two cases:

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

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

Notice that when $0 \leq (\boldsymbol{\mu}^* - \boldsymbol{\theta}'') \odot \mathbf{e}_{Z_2} \overset{(8.22)}{\leq} \varepsilon \mathbf{e}_{Z_2}$, then

$$\mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\Big) \geq \mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) - (m^* - 1)\varepsilon. \tag{8.23}$$

Indeed, (8.23) is a consequence of

$$\begin{aligned}
\mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c} - \boldsymbol{\theta}' \odot \mathbf{e}_{A^*} - \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) &= \mathbf{e}_{Z_2}^\mathsf{T}(\boldsymbol{\theta}'' - \boldsymbol{\theta}') \\
&= \mathbf{e}_{Z_2}^\mathsf{T}(\boldsymbol{\theta}'' - \boldsymbol{\mu}^*) + \mathbf{e}_{Z_2}^\mathsf{T}(\boldsymbol{\mu}^* - \boldsymbol{\theta}') \\
&\geq -\varepsilon(m^* - 1) + 0,
\end{aligned}$$

where we used (8.22), (8.15) and that Z_2 is strictly included in A^*.

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

$$\begin{aligned}
\mathbf{e}_B^\mathsf{T}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\Big) &\geq \mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\Big) \\
&\overset{(8.24)}{\geq} \mathbf{e}_A^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) - (m^* - 1)\varepsilon \\
&\geq \mathbf{e}_{A^*}^\mathsf{T}\Big(\boldsymbol{\theta}' \odot \mathbf{e}_{A^*} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{A^{*c}}\Big) - (m^* - 1)\varepsilon \\
&\overset{(8.25)}{\geq} \mathbf{e}_{A^*}^\mathsf{T}\boldsymbol{\mu}^* - m^*\varepsilon - (m^* - 1)\varepsilon,
\end{aligned}$$

where (8.24) uses (8.23) and (8.25) uses (8.16). This rewrites as

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

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 $0 \leq (\boldsymbol{\mu}^* - \boldsymbol{\theta}'') \odot \mathbf{e}_{Z_2} \leq \varepsilon \mathbf{e}_{Z_2}$ and an action $B \in \arg\max_{A' \in \mathcal{A}} \mathbf{e}_{A'}^\mathsf{T}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\Big)$ 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 $S' \neq \text{Oracle}\Big(\boldsymbol{\theta}'' \odot \mathbf{e}_{Z_2} + \widetilde{\boldsymbol{\theta}}_t \odot \mathbf{e}_{Z_2{}^c}\Big)$:

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

where the last inequality is obtained in the same way as in inequalities from (8.24) to (8.25).

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

□

Chapter 9

Conclusion and Perspectives

In this chapter, we conclude the thesis. In particular, we state the answers to the main research questions we raised, summarize the main contributions we have made and propose some directions for future research.

9.1 Conclusion

Let's remember that our main interest was to improve the efficiency of the policies, both from a statistical and computational point of view. We will therefore first look at how the thesis addressed this point.

Statistical point of view The statistical improvement of policies has been made on several levels in the thesis, which can be summarized in the following three aspects:

- **The improvement and extension of the regret analysis.** This point has been particularly addressed in Chapter 3, where we gave several results widely usable in the CMAB-T setting in order to prove regret upper bounds. Going further, a take-home message can be drawn from this chapter, namely that CMAB-T problems are essentially no more difficult than CMAB ones, because we can convert the probability $p_i(S_t)$ of observing the arm i into $\mathbb{I}\{i \in A_t\}$ using conditioning in the expectation of the regret. More specifically, in this chapter, we provided a simpler state-of-the-art ℓ_1 analysis under the ℓ_1-norm triggering probability modulated condition. In addition, under the same condition, we also provided several ℓ_2 analyses that are the first of this kind in CMAB-T. These last results opened the way to the use of ℓ_2 bonuses in CMAB-T problems, such as the OIM problem, that we considered in Chapter 6.

- **The improvement of confidence regions used in optimistic methods.** Apart from the improvement due to the shift from ℓ_∞ to ℓ_2 regions, the improvement of confidence regions was mainly explored in Chapters 4 and 7, where the use of auxiliary estimates helped to speed up the learning process of the mean. More precisely, in Chapter 4 for ℓ_∞ analysis, the use of the CUCB-V method improved the regret rate by a factor n, leveraging on the 1-sparsity of the outcomes. We extended this type of improvement in Chapter 7 for ℓ_2 analysis, using the estimation of the entire covariance matrix to refine the confidence regions. We notably proved a tight covariance-dependent regret bound for the resulting ESCB-C policy, outperforming over OLS-UCB. It should be noted in passing that the assumption we considered on the outcomes are more practical than the sub-Gaussian one. Finally, we were able to develop a tight approach to deal with the case of s-sparse outcomes.

- **The consideration of a non-optimistic policy.** We also considered randomized policies in Chapter 8, using a Thomspon sampling approach. We could see in particular that the regret rate of those policies were essentially the same as the ones for optimistic ℓ_2 policies. More precisely, we proposed CTS policies in two different settings, first with a beta prior for independent outcomes (which was already existing), and then with a Gaussian prior for sub-Gaussian outcomes. For the two cases, we gave a new tight regret analysis. We also noted that the empirical performance of CTS was generally superior compared to the optimistic ℓ_1 and ℓ_2 approaches.

Computational point of view Regarding the computational efficiency of policies, we can once again detail three aspects:

- **Exploitation of the uncertainty structure.** Although ℓ_2 approaches are more powerful, they pose some computational efficiency issues when used with optimistic policies. In Chapter 5, for CMAB with linear reward function, we were able to develop (for matroid-based action spaces) some methods to remedy the ℓ_2 inefficiency. More precisely, we gave several approximation algorithms for the maximization of a set function that is a sum of a linear one and a submodular one. This allowed us to give efficient implementations of the ESCB policy with matroid constraints. We also studied the budgeted setting, and gave an approximation scheme to adapt the above approach.

- **Finding an efficient surrogate for exploration bonuses.** In Chapter 6, we circumvented the inefficiency of the ℓ_2 methods by relaxing the exploration bonus under consideration, the goal being that the new bonus, although statistically less efficient, would be computationally efficient to use in the polities. More precisely, in the problems we looked at in this chapter (i.e., OIM and BOIM), the oracle (GREEDY) leverages the submodularity of the reward function to provide the $1 - 1/e$ approximation factor. We thus saved the computational efficiency by constructing submodular surrogate bonuses, since submodularity is closed under non-negative linear combinations.

- **Computational efficiency of the cts policy.** Faced with the difficulty of making optimistic ℓ_2 policies efficient, we proposed in Chapter 8 a non-optimistic alternative, namely CTS. As we have already seen, CTS obtained essentially the same performance as the ℓ_2 methods, while being computationally efficient to implement. Therefore, this approach is surely the best answer, in this thesis, to our initial challenge about the trade-off between the two efficiencies: it then seems that such a trade-off is, in fact, unnecessary. As we will see, we think future efforts should be made to try to transfer systematically any optimistic ℓ_2 approach to a CTS approach of the same performance.

Miscellaneous results in combinatorial optimization In this thesis, we also had the opportunity to study different combinatorial optimization problems. Indeed, in Chapter 4, we introduced a new sequential search-and-stop problem, that falls into the CMAB-T setting and the budgeted setting. The offline sequential problem led us to the formulation of an interesting new combinatorial optimization problem, which is a variant of the classical scheduling problem $1|prec|\sum w_j C_j$ where the agent can, whenever it wishes, restart its search all over again by re-shuffling the not found hidden object. We proposed an exact and efficient solution that makes an elegant link to the exact solution of the original $1|prec|\sum w_j C_j$ problem. The BOIM problem

considered in Chapter 6 also led us to formulate a new combinatorial optimization problem (in the same way, considering the budgeted version of an already known problem). This problem is the maximization of a ratio between a sub-modular function and a modular one. We contributed by improving the analysis of the GREEDY algorithm for this ratio maximization problem.

9.2 Perspectives

We have already mentioned some possible future works at the end of the corresponding chapters. Here, we detail those which seem to us the most relevant and promising with regard to our subject matter, i.e. those related to the CTS policy.

Correlated Gaussian prior for cts In our approach, even though we have treated the general case of correlated sub-Gaussian outcomes, the prior we used in the CTS policy is not (i.e., the components are independent Gaussian random variables). This independence has been very useful in the analysis, and an interesting open question is whether we can use a correlated prior. More specifically, under the sub-Gaussian assumption, we would like to use the prior

$$\mathcal{N}\left(\overline{\boldsymbol{\mu}}_{t-1}, \left(\frac{N_{ij,t-1}\Gamma_{ij}}{N_{i,t-1}N_{j,t-1}}\right)_{ij}\right), \tag{9.1}$$

which is tighter than the following that we used

$$\mathcal{N}\left(\overline{\boldsymbol{\mu}}_{t-1}, \operatorname{diag}\left(\max_{A'\in\mathcal{A},\ i\in A'} \sum_{j\in A'} \frac{\Gamma_{ij}}{N_{i,t-1}}\right)_i\right). \tag{9.2}$$

Apart from the gain in the empirical performance of the policy, there are some advantages to being able to conduct an analysis with the correlated prior (9.1). Indeed, it then allows us to conduct the Degenne and Perchet (2016b) analysis. This is a good point for the following reasons

- Admittedly, as we have seen, the bound obtained in the end would depend on the factor γ and is not as explicit as the one we have obtained (which depends on the whole covariance), but in cases close to the "arbitrarily correlated" case, the extra $\log^2(m)$ is removed (by the way, an interesting question is how to associate both a dependence in the whole covariance and a suppression of this $\log^2(m)$ factor when we approach the "arbitrarily correlated" case).
- We can still conduct the explicit analysis from (9.1), while returning to the Degenne and Perchet (2016b) analysis from (9.2) does not seem possible.
- This avoids having to calculate $\max_{A'\in\mathcal{A},\ i\in A'} \sum_{j\in A'} \frac{\Gamma_{ij}}{N_{i,t-1}}$, that may not be feasible efficiently. Actually, even if this is feasible, an extension that replaces $\boldsymbol{\Gamma}$ with a covariance estimation may raise some issues, as we will see in the next paragraph.

Covariance estimation for cts Another interesting question is whether an approach similar to the one presented in Chapter 7 can be used for CTS. There may be more than one way of doing this (e.g., with a prior or with an optimistic estimator on the covariance), but we argue here that the use of an independent prior as in

(9.2) may raise some issues. Indeed, the errors generated by the uncertainty that we have on the covariance estimates then concern non-observed coordinates (because of the maximum taken on $A' \in \mathcal{A}$). Having a bonus depending on the counters whose coordinates do not get feedback is redibitory for the analysis. In contrast, a prior of the form (9.1) doesn't have that inconvenience, and seems more promising to use for covariance estimation.

Approximation regret for cts Another unresolved difficulty for CTS is its use for approximation regrets. Although Wang and Chen (2018) have shown that this is generally impossible, exploiting the special design of some approximation oracles to prove that CTS works with them seems an interesting direction.

Bibliography

Abbasi-Yadkori, Yasin, David Pal, and Csaba Szepesvari (21–23 Apr 2012). "Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 1–9. URL: http://proceedings.mlr.press/v22/abbasi-yadkori12.html.

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved algorithms for linear stochastic bandits". In: *Neural Information Processing Systems*.

Abeille, Marc and Alessandro Lazaric (2017). "Linear Thompson sampling revisited". In: *International Conference on Artificial Intelligence and Statistics*. URL: http://proceedings.mlr.press/v54/abeille17a/abeille17a.pdf.

Abernethy, Jacob D, Elad Hazan, and Alexander Rakhlin (2008). "Competing in the dark: An efficient algorithm for bandit linear optimization." In: *Conference on Learning Theory*.

Agrawal, Rajeev (1995). "Sample mean based index policies by o (log n) regret for the multi-armed bandit problem". In: *Advances in Applied Probability* 27.4, pp. 1054–1078.

Agrawal, Shipra, Vashist Avadhanula, et al. (2017). "Thompson sampling for the mnl-bandit". In: *arXiv preprint arXiv:1706.00977*.

Agrawal, Shipra and Navin Goyal (2012a). "Analysis of Thompson sampling for the multi-armed bandit problem". In: *Conference on Learning Theory*.

— (Sept. 2012b). "Thompson Sampling for Contextual Bandits with Linear Payoffs". In: *CoRR, abs/1209.3352, http://arxiv.org/abs/1209.3352*. arXiv: 1209.3352. URL: http://arxiv.org/abs/1209.3352.

Alpern, Steve, Robbert Fokkink, et al. (2013). *Search theory*. Springer.

Alpern, Steve and Shmuel Gal (2006). *The theory of search games and rendezvous*. Vol. 55. Springer Science & Business Media.

Alpern, Steve and Thomas Lidbetter (2013). "Mining coal or finding terrorists: The expanding search paradigm". In: *Operations Research* 61.2, pp. 265–279.

Ambühl, Christoph and Monaldo Mastrolilli (2009). "Single machine precedence constrained scheduling is a vertex cover problem". In: *Algorithmica (New York)* 53.4, pp. 488–503. ISSN: 01784617. DOI: 10.1007/s00453-008-9251-6.

Ambühl, Christoph, Monaldo Mastrolilli, et al. (2011). "On the Approximability of Single-Machine Scheduling with Precedence Constraints". In: *Mathematics of Operations Research* 36.4, pp. 653–669. ISSN: 0364765X. DOI: 10.2307/41412330. URL: http://www.jstor.org/stable/41412330.

Anagnostopoulos, Aris et al. (2015). "Stochastic query covering for fast approximate document retrieval". In: *ACM Transactions on Information Systems (TOIS)* 33.3, pp. 1–35.

Anantharam, Venkatachalam, Pravin Varaiya, and Jean Walrand (1987). "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards". In: *IEEE Transactions on Automatic Control* 32.11, pp. 968–976.

Atamtürk, Alper and Andrés Gómez (2017). "Maximizing a class of utility functions over the vertices of a polytope". In: *Operations Research* 65.2, pp. 433–445.

Audibert, Jean Yves, Rémi Munos, and Csaba Szepesvári (2009a). "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits". In: *Theoretical Computer Science* 410.19, pp. 1876–1902. ISSN: 03043975. DOI: 10.1016/j.tcs.2009.01.016.

Audibert, Jean-Yves and Sébastien Bubeck (2010). "Regret Bounds and Minimax Policies under Partial Monitoring". In: *Journal of Machine Learning Research* 11, pp. 2785–2836.

Audibert, Jean-Yves, Sébastien Bubeck, and Gabor Lugosi (2011). "Minimax Policies for Combinatorial Prediction Games". In: *Proceedings of the 24th annual Conference On Learning Theory*. COLT '11.

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

Auer, Peter (2002). "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3, pp. 397–422.

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer (2002). "Finite-time analysis of the multiarmed bandit problem". In: *Machine Learning* 47.2-3, pp. 235–256.

Aziz, Maryam, Emilie Kaufmann, and Marie-Karelle Riviere (2019). "On Multi-Armed Bandit Designs for Phase I Clinical Trials". In: *arXiv preprint arXiv:1903.07082*.

Babaioff, Moshe, Yogeshwer Sharma, and Aleksandrs Slivkins (2009). "Characterizing truthful multi-armed bandit mechanisms". In: *ACM-EC*, pp. 79–88.

Bach, Francis (2011). "Learning with Submodular Functions: A Convex Optimization Perspective". In: arXiv: 1111.6453. URL: http://arxiv.org/abs/1111.6453.

Badanidiyuru, Ashwinkumar, Robert Kleinberg, and Aleksandrs Slivkins (2013). "Bandits with knapsacks". In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 207–216. ISBN: 9780769551357. DOI: 10.1109/FOCS.2013.30. arXiv: 1305.2545.

Bai, Wenruo et al. (2016). "Algorithms for optimizing the ratio of submodular functions". In: *International Conference on Machine Learning*, pp. 2751–2759.

Bernstein, Sergei (1924). "On a modification of Chebyshev's inequality and of the error formula of Laplace". In: *Ann. Sci. Inst. Sav. Ukraine, Sect. Math* 1.4, pp. 38–49.

Berry, Donald A and Bert Fristedt (1985). *Bandit Problems: Sequential Allocation of Experiments*. Vol. 38. Monographs on statistics and applied probability 8. Chapman and Hall, pp. viii, 275.

Besson, Lilian and Emilie Kaufmann (2017). "Multi-player bandits revisited". In: *arXiv preprint arXiv:1711.02317*.

Borcea, Julius, Petter Brändén, and Thomas Liggett (2009). "Negative dependence and the geometry of polynomials". In: *Journal of the American Mathematical Society* 22.2, pp. 521–567.

Boursier, Etienne and Vianney Perchet (2019). "SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits". In: *Advances in Neural Information Processing Systems*, pp. 12048–12057.

Browne, Cameron B et al. (2012). "A survey of Monte Carlo tree search methods". In: *IEEE Transactions on Computational Intelligence and AI in Games* 4.1, pp. 1–43.

Brualdi, Richard A. (1969). "Comments on bases in dependence structures". In: *Bulletin of the Australian Mathematical Society* 1.2, pp. 161–167. ISSN: 17551633. DOI: 10.1017/S000497270004140X.

Bubeck, Sébastien and Nicolò Cesa-Bianchi (2012). "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5, pp. 1–122. arXiv: 1204.5721. URL: http://arxiv.org/abs/1204.5721.

Bubeck, Sébastien, Nicolò Cesa-Bianchi, and Sham M Kakade (2012). "Towards minimax policies for online linear optimization with bandit feedback". In: *Conference on Learning Theory*.

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

Bubeck, Sébastien, Damien Ernst, and Aurélien Garivier (2013). "Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality". In: *Journal of Machine Learning Research* 14, pp. 601–623. URL: http://www.jmlr.org/papers/volume14/bubeck13a/bubeck13a.pdf.

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

Buldygin, Valerii V and Yu V Kozachenko (1980). "Sub-Gaussian random variables". In: *Ukrainian Mathematical Journal* 32.6, pp. 483–489.

Burnetas, Apostolos N and Michael N Katehakis (1996). "Optimal adaptive policies for sequential allocation problems". In: *Advances in Applied Mathematics* 17.2, pp. 122–142.

Cappé, Olivier et al. (2013). "Kullback–leibler upper confidence bounds for optimal sequential allocation". In: *The Annals of Statistics* 41.3, pp. 1516–1541.

Caro, Felipe and Jérémie Gallien (2007). "Dynamic assortment with demand learning for seasonal consumer goods". In: *Management Science* 53.2, pp. 276–292.

Carpentier, Alexandra and Remi Munos (21–23 Apr 2012). "Bandit Theory meets Compressed Sensing for high dimensional Stochastic Linear Bandit". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 190–198. URL: http://proceedings.mlr.press/v22/carpentier12.html.

Carpentier, Alexandra and Michal Valko (2016). "Revealing graph bandits for maximizing local influence". In: *International Conference on Artificial Intelligence and Statistics*.

Cesa-Bianchi, Nicolò and Paul Fischer (1998). "Finite-Time Regret Bounds for the Multiarmed Bandit Problem." In: *ICML*. Vol. 1998. Citeseer, pp. 100–108.

Cesa-Bianchi, Nicolò, Claudio Gentile, and Giovanni Zappella (2013). "A gang of bandits". In: *Neural Information Processing Systems*.

Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.

— (2012). "Combinatorial bandits". In: *Journal of Computer and System Sciences*. Vol. 78. 5, pp. 1404–1422.

Chang, Seok-Ho, Pamela C Cosman, and Laurence B Milstein (2011). "Chernoff-type bounds for the Gaussian error function". In: *IEEE Transactions on Communications* 59.11, pp. 2939–2944.

Chapelle, Olivier and Lihong Li (2011). "An empirical evaluation of Thompson sampling". In: *Neural Information Processing Systems*. URL: https://arxiv.org/pdf/1605.08722.pdf.

Chen, Wei, Chi Wang, and Yajun Wang (2010). "Scalable influence maximization for prevalent viral marketing in large-scale social networks". In: *Knowledge Discovery and Data Mining*.

Chen, Wei, Yajun Wang, and Yang Yuan (2013). "Combinatorial Multi-Armed Bandit: General Framework and Applications". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR, pp. 151–159. URL: http://proceedings.mlr.press/v28/chen13a.html.

— (2016). "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms". In: *Journal of Machine Learning Research* 17.

Combes, Richard et al. (2015). "Combinatorial Bandits Revisited". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., pp. 2116–2124. URL: http://papers.nips.cc/paper/5831-combinatorial-bandits-revisited.pdf.

Coquelin, Pierre-Arnaud and Rémi Munos (2007). "Bandit algorithms for tree search". In: *Uncertainty in Artificial Intelligence*.

Coulom, Rémi (2007). "Efficient selectivity and backup operators in Monte-Carlo tree search". In: *Computers and games* 4630, pp. 72–83.

Cuvelier, Thibaut, Richard Combes, and Eric Gourdin (2020). "Statistically Efficient, Polynomial Time Algorithms for Combinatorial Semi Bandits". In: *arXiv preprint arXiv:2002.07258*.

Dani, Varsha, Thomas Hayes, and Sham Kakade (2008). "The Price of Bandit Information for Online Optimization". In: *Advances in Neural Information Processing Systems 20* 20. Ed. by J C Platt et al., pp. 1–8. ISSN: 00368075. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.4607%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

Degenne, Rémy and Vianney Perchet (20–22 Jun 2016a). "Anytime optimal algorithms in stochastic multi-armed bandits". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1587–1595. URL: http://proceedings.mlr.press/v48/degenne16.html.

— (2016b). "Combinatorial semi-bandit with known covariance". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., pp. 2972–2980. URL: http://papers.nips.cc/paper/6137-combinatorial-semi-bandit-with-known-covariance.pdf.

Dijkstra, Edsger W et al. (1959). "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1, pp. 269–271.

Ding, Wenkui, Tao Qin, et al. (2013). "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. ISBN: 9781577356158. URL: http://dblp.uni-trier.de/db/conf/aaai/aaai2013.html%7B%5C#%7DDingQZL13.

Ding, Wenkui, Tao Qiny, et al. (2013). *Multi-armed bandit with budget constraint and variable costs*. URL: https://dl.acm.org/citation.cfm?id=2891493%7B%5C#%7D.W5fpf5BQ7RY.mendeley.

Durrett, Rick (2019). *Probability: theory and examples*. Vol. 49. Cambridge university press.

Edmonds, Jack (1971). "Matroids and the Greedy Algorithm". In: *Mathematical Programming* 1.1, pp. 127–136.

Edmonds, Jack and D.R. Fulkerson (1965). "Transversals and matroid partition". In: *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics* 69B.3, pp. 147–153. ISSN: 0022-4340. DOI: 10.6028/jres.069B.016. URL: http://nvlpubs.nist.gov/nistpubs/jres/69B/jresv69Bn3p147%7B%5C_%7DA1b.pdf.

Evans, Thomas P Oléron and Steven R Bishop (2013). "Static search games played over graphs and general metric spaces". In: *European Journal of Operational Research* 231.3, pp. 667–689.

Feige, Uriel (1998). "A threshold of ln n for approximating set cover". In: *Journal of the ACM (JACM)* 45.4, pp. 634–652.

Filmus, Yuval and Justin Ward (2012). "A tight combinatorial algorithm for submodular maximization subject to a matroid constraint". In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 659–668. ISBN: 978-0-7695-4874-6. DOI: 10.1109/FOCS.2012.55. arXiv: arXiv:1204.4526v2.

Flajolet, Arthur and Patrick Jaillet (Oct. 2015). "Logarithmic regret bounds for Bandits with Knapsacks". In: arXiv: 1510.01800. URL: http://arxiv.org/abs/1510.01800.

Fokkink, Robbert, Thomas Lidbetter, and László A. Végh (July 2016). "On Submodular Search and Machine Scheduling". In: arXiv: 1607.07598. URL: http://arxiv.org/abs/1607.07598.

Fujishige, Satoru (2005). *Submodular functions and optimization*. Vol. 58. Elsevier.

Gagliolo, Matteo and Jürgen Schmidhuber (2010). "Algorithm selection as a bandit problem with unbounded losses". In: *International Conference on Learning and Intelligent Optimization*. Springer, pp. 82–96.

Gai, Yi, Bhaskar Krishnamachari, and Rahul Jain (2010). "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)*. IEEE, pp. 1–9.

— (2012). "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations". In: *Transactions on Networking* 20.5, pp. 1466–1478.

Gal, Shmuel (2001). "On the optimality of a simple strategy for searching graphs". In: *International Journal of Game Theory* 29.4, pp. 533–542.

Garivier, Aurélien and Olivier Cappé (2011). "The {KL}-{UCB} algorithm for bounded stochastic bandits and beyond". In: *Proceedings of the 24th annual Conference On Learning Theory*. COLT '11.

Gelly, Sylvain et al. (2006). *Modification of UCT with patterns in Monte-Carlo Go*. Tech. rep. Inria. URL: https://hal.inria.fr/inria-00117266.

Gentile, Claudio, Shuai Li, and Giovanni Zappella (2014). "Online clustering of bandits". In: *International Conference on Machine Learning*.

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

Gittins, John C, Richard Weber, and Kevin Glazebrook (1989). *Multi-armed Bandit Allocation Indices*. Wiley.

Goemans, Michel X et al. (2009). "Approximating submodular functions everywhere". In: *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, pp. 535–544.

Gomez-Rodriguez, Manuel, Jure Leskovec, and Andreas Krause (Feb. 2012). "Inferring Networks of Diffusion and Influence". In: *ACM Trans. Knowl. Discov. Data*

5.4, 21:1–21:37. ISSN: 1556-4681. DOI: 10.1145/2086737.2086741. URL: http://doi.acm.org/10.1145/2086737.2086741.

Gopalan, Aditya, Shie Mannor, and Yishay Mansour (2014). "Thompson sampling for complex bandit problems". In: *International Conference on Machine Learning*.

Goyal, A., W. Lu, and L. V. S. Lakshmanan (Dec. 2011). "SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model". In: *2011 IEEE 11th International Conference on Data Mining*, pp. 211–220. DOI: 10.1109/ICDM.2011.132.

Goyal, Amit, Francesco Bonchi, and Laks V.S. Lakshmanan (2010). "Learning Influence Probabilities in Social Networks". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. WSDM '10. New York, New York, USA: ACM, pp. 241–250. ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718518. URL: http://doi.acm.org/10.1145/1718487.1718518.

Graham, R. L. et al. (1979). "Optimization and approximation in deterministic sequencing and scheduling: A survey". In: *Annals of Discrete Mathematics* 5.C, pp. 287–326. ISSN: 01675060. DOI: 10.1016/S0167-5060(08)70356-X.

Graves, Todd L and Tze Leung Lai (1997). "Asymptotically efficient adaptive choice of control laws incontrolled markov chains". In: *SIAM journal on control and optimization* 35.3, pp. 715–743.

Gupta, Anupam and Viswanath Nagarajan (2013). "A stochastic probing problem with applications". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7801 LNCS, pp. 205–216. ISBN: 9783642366932. DOI: 10.1007/978-3-642-36694-9_18. arXiv: arXiv:1302.5913v1.

György, András et al. (2007). "The on-line shortest path problem under partial monitoring". In: *Journal of Machine Learning Research* 8.Oct, pp. 2369–2403.

He, Simai, Jiawei Zhang, and Shuzhong Zhang (2012). "Polymatroid Optimization, Submodularity, and Joint Replenishment Games". In: *Operations Research* 60.1, pp. 128–137. ISSN: 0030-364X. DOI: 10.1287/opre.1110.1000. URL: http://pubsonline.informs.org/doi/abs/10.1287/opre.1110.1000.

He, Zengyou, Xiaofei Xu, and Shengchun Deng (Mar. 2006). "Mining top-k strongly correlated item pairs without minimum correlation threshold". In: *KES Journal* 10, pp. 105–112. DOI: 10.3233/KES-2006-10202.

Heckerman, David, John S. Breese, and Koos Rommelse (1995). "Decision-theoretic troubleshooting". In: *Communications of the ACM* 38.3, pp. 49–57. ISSN: 00010782. DOI: 10.1145/203330.203341. URL: http://portal.acm.org/citation.cfm?doid=203330.203341.

Hochbaum, Dorit S (1996). "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems". In: *Approximation algorithms for NP-hard problems*, pp. 94–143.

— (1997). "Approximation algorithms for NP-hard problems". In: *ACM Sigact News* 28.2, pp. 40–52.

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

Hoffman, Matthew D, Eric Brochu, and Nando de Freitas (2011). "Portfolio Allocation for Bayesian Optimization." In: *UAI*. Citeseer, pp. 327–336.

Hohzaki, Ryusuke (2016). "Search games: Literature and survey". In: *Journal of the Operations Research Society of Japan* 59.1, pp. 1–34.

Honorio, Jean and Tommi Jaakkola (22–25 Apr 2014). "Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence*

and Statistics. Ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, pp. 384–392. URL: http://proceedings.mlr.press/v33/honorio14.html.

Horn, Roger A and Charles R Johnson (1990). *Matrix analysis*. Cambridge University Press.

Jacobs, Irwin Mark and JM Wozencraft (1965). "Principles of communication engineering." In:

Jensen, Finn V. et al. (2001). "The SACSO methodology for troubleshooting complex systems". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM* 15.4, pp. 321–333. ISSN: 08900604. DOI: 10.1017/S0890060401154065.

Jukna, Stasys and Georg Schnitger (2016). "On the optimality of Bellman–Ford–Moore shortest path algorithm". In: *Theoretical Computer Science* 628, pp. 101–109.

Kaggle (2013). URL: https://www.kaggle.com/.

Kakade, Sham M, Adam Tauman Kalai, and Katrina Ligett (2009). "Playing games with approximation algorithms". In: *SIAM Journal on Computing* 39.3, pp. 1088–1106.

Kaufmann, Emilie, Nathaniel Korda, and Rémi Munos (2012). "Thompson Sampling: An Asymptotically Optimal Finite Time Analysis". In: *Algorithmic Learning Theory*.

Kempe, David, Jon Kleinberg, and Éva Tardos (2015). "Maximizing the spread of influence through a social network". In: *Theory of Computing* 11.4, pp. 105–147.

Khuller, Samir, Anna Moss, and Joseph Seffi Naor (1999). "The budgeted maximum coverage problem". In: *Information processing letters* 70.1, pp. 39–45.

Kikuta, Kensaku and William H Ruckle (1994). "Initial point search on weighted trees". In: *Naval Research Logistics (NRL)* 41.6, pp. 821–831.

Kocsis, Levente and Csaba Szepesvári (2006). "Bandit-based Monte-Carlo planning". In: *European Conference on Machine Learning*.

Komiyama, Junpei, Junya Honda, and Hiroshi Nakagawa (June 2015). "Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays". In: arXiv: 1506.00779. URL: http://arxiv.org/abs/1506.00779.

Krause, Andreas and Daniel Golovin (2011). "Submodular function maximization". In: *Tractability*. Vol. 9781107025, pp. 71–104. ISBN: 9781139177801. DOI: 10.1017/CBO9781139177801.004.

Krause, Andreas and Carlos Guestrin (2005). *A Note on the Budgeted Maximization of Submodular Functions*. Tech. rep. June. CMU, pp. 1–7. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.9721%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

— (2007). "Near-optimal observation selection using submodular functions". In: *AAAI*. Vol. 7, pp. 1650–1654.

Kuhn, Harold W (1955). "The Hungarian method for the assignment problem". In: *Naval research logistics quarterly* 2.1-2, pp. 83–97.

Kullback, Solomon and Richard A Leibler (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Kveton, Branislav, Csaba Szepesvari, et al. (2015). "Cascading Bandits: Learning to Rank in the Cascade Model". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 767–776. URL: http://proceedings.mlr.press/v37/kveton15.html.

Kveton, Branislav, Zheng Wen, Azin Ashkan, Hoda Eydgahi, et al. (2014). "Matroid bandits: Fast combinatorial optimization with learning". In: *Uncertainty in Artificial Intelligence*.

Kveton, Branislav, Zheng Wen, Azin Ashkan, and Csaba Szepesvari (July 2015a). "Combinatorial Cascading Bandits". In: arXiv: 1507.04208. URL: http://arxiv.org/abs/1507.04208.

— (2015b). "Tight regret bounds for stochastic combinatorial semi-bandits". In: *International Conference on Artificial Intelligence and Statistics*.

Kwon, Joon and Vianney Perchet (2015). "Gains and Losses are Fundamentally Different in Regret Minimization: The Sparse Case". In: *CoRR* abs/1511.08405. arXiv: 1511.08405. URL: http://arxiv.org/abs/1511.08405.

Kwon, Joon, Vianney Perchet, and Claire Vernade (2017). "Sparse Stochastic Bandits". In: *CoRR* abs/1706.01383. arXiv: 1706.01383. URL: http://arxiv.org/abs/1706.01383.

Lagrée, Paul, Claire Vernade, and Olivier Cappe (2016). "Multiple-play bandits in the position-based model". In: *Advances in Neural Information Processing Systems*, pp. 1597–1605.

Lai, Lifeng, Hai Jiang, and H Vincent Poor (2008). "Medium access in cognitive radio networks: A competitive multi-armed bandit framework". In: *2008 42nd Asilomar Conference on Signals, Systems and Computers*. IEEE, pp. 98–102.

Lai, Tze L and Herbert Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: *Advances in Applied Mathematics* 6.1, pp. 4–22.

Lattimore, Tor and Csaba Szepesvári (2019). *Bandit algorithms*. URL: http://downloads.tor-lattimore.com/book.pdf.

Laurent, Beatrice and Pascal Massart (2000). "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics*, pp. 1302–1338.

Lawler, E. L. (1978). "Sequencing jobs to minimize total weighted completion time subject to precedence constraints". In: *Annals of Discrete Mathematics* 2.C, pp. 75–90. ISSN: 01675060. DOI: 10.1016/S0167-5060(08)70323-6.

Lee, Jon et al. (2010). "Maximizing Nonmonotone Submodular Functions under Matroid or Knapsack Constraints". In: *SIAM Journal on Discrete Mathematics* 23.4, pp. 2053–2078. ISSN: 0895-4801. DOI: 10.1137/090750020. arXiv: 0902.0353. URL: http://epubs.siam.org/doi/10.1137/090750020.

Lenstra, J. K. and A. H. G. Rinnooy Kan (1978). "Complexity of Scheduling under Precedence Constraints". In: *Operations Research* 26.1, pp. 22–35. ISSN: 0030-364X. DOI: 10.1287/opre.26.1.22. URL: http://pubsonline.informs.org/doi/abs/10.1287/opre.26.1.22.

Leskovec, Jure, Andreas Krause, et al. (2007). "Cost-effective Outbreak Detection in Networks". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: ACM, pp. 420–429. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281239. URL: http://doi.acm.org/10.1145/1281192.1281239.

Leskovec, Jure and Andrej Krevl (June 2014). *{SNAP Datasets}: {Stanford} Large Network Dataset Collection*. \url{http://snap.stanford.edu/data}.

Li, Lihong, Wei Chu, et al. (2010). "A contextual-bandit approach to personalized news article recommendation". In: *International World Wide Web Conference*.

Li, Shuai, Alexandros Karatzoglou, and Claudio Gentile (2016). "Collaborative Filtering Bandits". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: ACM, pp. 539–548. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2911548. URL: http://doi.acm.org/10.1145/2911451.2911548.

Li, Shuai, Baoxiang Wang, et al. (2016). "Contextual Combinatorial Cascading Bandits". In: *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1245–1253.

Lín, Václav (2015). "Scheduling results applicable to decision-theoretic troubleshooting". In: *International Journal of Approximate Reasoning* 56.PA, pp. 87–107. ISSN: 0888613X. DOI: 10.1016/j.ijar.2014.08.004.

Liu, Haoyang, Keqin Liu, and Qing Zhao (2011). "Logarithmic weak regret of non-Bayesian restless multi-armed bandit". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1968–1971.

Liu, Keqin and Qing Zhao (2012). "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)*. IEEE, pp. 232–237.

Lovász, László (1983). "Submodular functions and convexity". In: *Mathematical programming the state of the art*. Ed. by Armin Bachem, Martin Grötschel, and Bernhard H Korte, pp. 235–257. URL: http://www.cs.elte.hu/%7B%5C%%7D7B%7B~%7D%7B%5C%%7D7Dlovasz/scans/submodular.pdf.

Lueker, George S (1975). *Two NP-complete problems in nonnegative integer programming*. Princeton University. Department of Electrical Engineering.

Lugosi, Gábor and Abbas Mehrabian (2018). "Multiplayer bandits without observing collision information". In: *arXiv preprint arXiv:1808.08416*.

Lugosi, Gábor, Gergely Neu, and Julia Olkhovskaya (22–24 Mar 2019). "Online Influence Maximization with Local Observations". In: *Proceedings of the 30th International Conference on Algorithmic Learning Theory*. Vol. 98. Proceedings of Machine Learning Research. Chicago, Illinois: PMLR, pp. 557–580. URL: http://proceedings.mlr.press/v98/lugosi19a.html.

Madhawa, Kaushalya and Tsuyoshi Murata (2019). "A multi-armed bandit approach for exploring partially observed networks". In: *Applied Network Science* 4.1, p. 26.

Magureanu, Stefan, Richard Combes, and Alexandre Proutiere (13–15 Jun 2014). "Lipschitz Bandits: Regret Lower Bound and Optimal Algorithms". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 975–999. URL: http://proceedings.mlr.press/v35/magureanu14.html.

Maillard, Odalric-Ambrym, Rémi Munos, and Gilles Stoltz (2011). "Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences". In: *To appear in Proceedings of the 24th annual Conference On Learning Theory*. COLT '11.

Mannor, Shie and Ohad Shamir (2011). "From bandits to experts: On the value of side-observations". In: *Advances in Neural Information Processing Systems*, pp. 684–692.

Marchal, Olivier, Julyan Arbel, et al. (2017). "On the sub-Gaussianity of the Beta and Dirichlet distributions". In: *Electronic Communications in Probability* 22.

McGregor, Andrew and Hoa T Vu (2019). "Better streaming algorithms for the maximum coverage problem". In: *Theory of Computing Systems* 63.7, pp. 1595–1619.

Merlis, Nadav and Shie Mannor (May 2019). "Batch-Size Independent Regret Bounds for the Combinatorial Multi-Armed Bandit Problem". In: arXiv: 1905.03125. URL: https://arxiv.org/abs/1905.03125.

Minoux, M (1978). "Accelerated greedy algorithms for maximizing submodular set functions". In: *Optimization Techniques*, pp. 234–243.

Mitzenmacher, Michael and Eli Upfal (2017). *Probability and computing: randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.

Mohri, Mehryar and Andres Munoz (2014). "Optimal regret minimization in posted-price auctions with strategic buyers". In: *Advances in Neural Information Processing Systems*, pp. 1871–1879.

Mukherjee, Subhojyoti et al. (Nov. 2017). "Efficient-UCBV: An Almost Optimal Algorithm using Variance Estimates". In: arXiv: 1711.03591. URL: https://arxiv.org/abs/1711.03591.

Nemhauser, G L, L A Wolsey, and M L Fisher (1978). "An analysis of approximations for maximizing submodular set functions–I". In: *Mathematical Programming* 14.1, pp. 265–294.

Netrapalli, Praneeth and Sujay Sanghavi (June 2012). "Learning the Graph of Epidemic Cascades". In: *SIGMETRICS Perform. Eval. Rev.* 40.1, pp. 211–222. ISSN: 0163-5999. DOI: 10.1145/2318857.2254783. URL: http://doi.acm.org/10.1145/2318857.2254783.

Neu, Gergely and Gábor Bartók (2013). "An efficient algorithm for learning with semi-bandit feedback". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8139 LNAI, pp. 234–248. ISBN: 9783642409349. DOI: 10.1007/978-3-642-40935-6_17. arXiv: 1305.2732.

Nguyen, Huy and Rong Zheng (2013). "On budgeted influence maximization in social networks". In: *IEEE Journal on Selected Areas in Communications* 31.6, pp. 1084–1094.

Nie, Xinkun et al. (2017). "Why adaptively collected data have negative bias and how to correct for it". In: *arXiv preprint arXiv:1708.01977*.

Oliveira, Carlos A.S. and Panos M. Pardalos (2005). *A survey of combinatorial optimization problems in multicast routing*. DOI: 10.1016/j.cor.2003.12.007.

Ontanón, Santiago (2013). "The combinatorial multi-armed bandit problem and its application to real-time strategy games". In: *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Pandey, Sandeep and Christopher Olston (2007). "Handling Advertisements of Unknown Quality in Search Advertising". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, pp. 1065–1072. URL: http://papers.nips.cc/paper/3012-handling-advertisements-of-unknown-quality-in-search-advertising.pdf.

Pathan, Mukaddim, Rajkumar Buyya, and Athena Vakali (2008). "Content delivery networks: State of the art, insights, and imperatives". In: *Content Delivery Networks*. Springer, pp. 3–32.

Peña, Victor H, Tze Leung Lai, and Qi-Man Shao (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Perfect, Hazel (1968). "Applications of Menger's graph theorem". In: *Journal of Mathematical Analysis and Applications* 22.1, pp. 96–111. ISSN: 10960813. DOI: 10.1016/0022-247X(68)90163-7.

Perrault, Pierre, Etienne Boursier, et al. (2020). "Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits". In: *arXiv preprint arXiv:2006.06613*.

Perrault, Pierre, Jennifer Healey, et al. (2020a). "Budgeted Online Influence Maximization". In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 6588–6599.

— (2020b). *On the Approximation Relationship between Optimizing Ratio of Submodular (RS) and Difference of Submodular (DS) Functions*.

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

— (2019b). "Finding the bandit in a graph: Sequential search-and-stop". In: *Proceedings of Machine Learning Research*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1668–1677. URL: http://proceedings.mlr.press/v89/perrault19a.html.

— (2020). "Covariance-adapting algorithm for semi-bandits with application to sparse rewards". In: *Conference on Learning Theory*.

Prot, D. and O. Bellenguez-Morineau (2017). *A survey on how the structure of precedence constraints may change the complexity class of scheduling problems*. DOI: 10.1007/s10951-017-0519-z. arXiv: 1510.04833.

Radlinski, Filip, Robert Kleinberg, and Thorsten Joachims (2008). "Learning diverse rankings with multi-armed bandits". In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 784–791.

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

Rosenski, Jonathan, Ohad Shamir, and Liran Szlak (2016). "Multi-player bandits–a musical chairs approach". In: *International Conference on Machine Learning*, pp. 155–163.

Rossi, Ryan A. and Nesreen K. Ahmed (2015). "The Network Data Repository with Interactive Graph Analytics and Visualization". In: *AAAI*. URL: http://networkrepository.com.

Russo, Daniel and Benjamin Van Roy (2016). "An information-theoretic analysis of thompson sampling". In: *The Journal of Machine Learning Research* 17.1, pp. 2442–2471.

Saha, Barna and Lise Getoor (2009). "On maximum coverage in the streaming model & application to multi-topic blog-watch". In: *Proceedings of the 2009 siam international conference on data mining*. SIAM, pp. 697–708.

Saito, Kazumi, Ryohei Nakano, and Masahiro Kimura (2008). "Prediction of Information Diffusion Probabilities for Independent Cascade Model". In: *Knowledge-Based Intelligent Information and Engineering Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 67–75. ISBN: 978-3-540-85567-5.

Sankararaman, Karthik Abinav and Aleksandrs Slivkins (May 2017). "Combinatorial Semi-Bandits with Knapsacks". In: arXiv: 1705.08110. URL: http://arxiv.org/abs/1705.08110.

Sauré, Denis and Assaf Zeevi (2013). "Optimal dynamic assortment planning with demand learning". In: *Manufacturing & Service Operations Management* 15.3, pp. 387–404.

Schrijver, Alexander (2008). *Combinatorial Optimization: Polyhedra and Efficiency*, p. 1920. ISBN: 3540204563. DOI: 10.1007/978-3-642-24508-4. arXiv: arXiv:1011.1669v3. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Algorithms+and+Combinatorics+24%7B%5C#%7D9.

Shaked, Moshe and J George Shanthikumar (2007). *Stochastic orders*. Springer Science & Business Media.

Sidney, J. B. (1975). "Decomposition Algorithms for Single-Machine Sequencing with Precedence Relations and Deferral Costs". In: *Operations Research* 23.2, pp. 283–

298. ISSN: 0030-364X. DOI: 10.1287/opre.23.2.283. URL: http://or.journal.informs.org/cgi/doi/10.1287/opre.23.2.283.

Silver, David, Aja Huang, et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489.

Silver, David, Julian Schrittwieser, et al. (2017). "Mastering the game of go without human knowledge". In: *nature* 550.7676, pp. 354–359.

Smith, Wayne E (1956). "Various optimizers for single-stage production". In: *Naval Research Logistics (NRL)* 3.1-2, pp. 59–66.

Stobbe, Peter and Andreas Krause (Oct. 2010). "Efficient Minimization of Decomposable Submodular Functions". In: arXiv: 1010.5511. URL: http://arxiv.org/abs/1010.5511.

Stone, Lawrence D (1976). *Theory of optimal search*. Vol. 118. Elsevier.

Streeter, Matthew and Daniel Golovin (2009). "An online algorithm for maximizing submodular functions". In: *Advances in Neural Information Processing Systems*, pp. 1577–1584.

Sutton, Richard and Andrew Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sviridenko, Maxim, Jan Vondrák, and Justin Ward (Nov. 2013). "Optimal approximation for submodular and supermodular optimization with bounded curvature". In: arXiv: 1311.4728. URL: http://arxiv.org/abs/1311.4728.

Talebi, M. Sadegh, Zhenhua Zou, et al. (Sept. 2013). "Stochastic Online Shortest Path Routing: The Value of Feedback". In: arXiv: 1309.7367. URL: https://arxiv.org/abs/1309.7367.

Talebi, Mohammad Sadegh and Alexandre Proutiere (2016). "An Optimal Algorithm for Stochastic Matroid Bandit Optimization". In: *The 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 548–556. ISBN: 9781450342391.

Tang, Youze, Yanchen Shi, and Xiaokui Xiao (2015). "Influence Maximization in Near-Linear Time: A Martingale Approach". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: ACM, pp. 1539–1554. ISBN: 978-1-4503-2758-9. DOI: 10.1145/2723372.2723734. URL: http://doi.acm.org/10.1145/2723372.2723734.

Tang, Youze, Xiaokui Xiao, and Yanchen Shi (2014). "Influence maximization: Near-optimal time complexity meets practical efficiency". In: *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, pp. 75–86.

Thompson, William R (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25, pp. 285–294.

— (1935). "On the theory of apportionment". In: *American Journal of Mathematics* 57.2, pp. 450–456.

Tran-Thanh, Long, Archie C Chapman, et al. (2012). *Knapsack Based Optimal Policies for Budget-Limited Multi-Armed Bandits*.

Tran-Thanh, Long, Sebastian Stein, et al. (2014). "Efficient crowdsourcing of unknown experts using bounded multi-armed bandits". In: *Artificial Intelligence* 214, pp. 89–111.

Tsybakov, Alexandre B (2009). *Springer Series in Statistics*.

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

Valko, Michal et al. (2014). "Spectral bandits for smooth graph functions". In: *International Conference on Machine Learning*.

Vanchinathan, Hastagiri P et al. (2015). "Discovering valuable items from massive data". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1195–1204.

Vaswani, Sharan, Branislav Kveton, et al. (2017). "Model-independent online learning for influence maximization". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 3530–3539.

Vaswani, Sharan, Laks. V S Lakshmanan, and Mark Schmidt (2015). "Influence maximization with bandits". In: *NIPS workshop on Networks in the Social and Information Sciences 2015*.

Vershynin, Roman (2009). *A note on sums of independent random matrices after Ahlswede-Winter*. Tech. rep. URL: http://www.umich.edu/%7B%5C%%7D7B%7B~%7D%7B%5C%%7D7Dromanv/teaching/reading-group/ahlswede-winter.pdf.

Villar, Sofia S, Jack Bowden, and James Wason (2015). "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 30.2, p. 199.

Vogel, Walter (1960). "An asymptotic minimax theorem for the two armed bandit problem". In: *The Annals of Mathematical Statistics* 31.2, pp. 444–451.

Wang, D., B. Song, et al. (June 2019). "Intelligent Cognitive Radio in 5G: AI-Based Hierarchical Cognitive Cellular Networks". In: *IEEE Wireless Communications* 26.3, pp. 54–61. ISSN: 1558-0687. DOI: 10.1109/MWC.2019.1800353.

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

Wang, Shatian, Shuoguang Yang, et al. (2020). "Fast Thompson Sampling Algorithm with Cumulative Oversampling: Application to Budgeted Influence Maximization". In: *arXiv preprint arXiv:2004.11963*.

Wang, Siwei and Wei Chen (Mar. 2018). "Thompson Sampling for Combinatorial Semi-Bandits". In: arXiv: 1803.04623. URL: http://arxiv.org/abs/1803.04623.

Wang, Yingfei, Hua Ouyang, et al. (1997). "Thompson Sampling for Contextual Combinatorial Bandits". In: *WOODSTOK '97*. DOI: 10.475/123. URL: http://asamov.com/download/ts%7B%5C_%7Dcombinatorial.pdf.

Wang, Yizao and Sylvain Gelly (2007). "Modifications of UCT and sequence-like simulations for Monte-Carlo Go". In: *2007 IEEE Symposium on Computational Intelligence and Games*. IEEE, pp. 175–182.

Watanabe, Ryo et al. (Nov. 2017). "KL-UCB-Based Policy for Budgeted Multi-Armed Bandits with Stochastic Action Costs". In: E100.A, pp. 2470–2486.

Watkins, Christopher John Cornish Hellaby (1989). "Learning from Delayed Rewards". In: *King's College, Cambridge*.

Wen, Zheng, Branislav Kveton, and Azin Ashkan (2015). "Efficient learning in large-scale combinatorial semi-bandits". In: *International Conference on Machine Learning*, pp. 1113–1122.

Wen, Zheng, Branislav Kveton, Michal Valko, et al. (2017). "Online influence maximization under independent cascade model with semi-bandit feedback". In: *Neural Information Processing Systems*.

Whitney, Hassler (1935). "On the abstract properties of linear dependence". In: *American Journal of Mathematics* 57.3, pp. 509–533.

Xia, Yingce, Wenkui Ding, et al. (2016). "Budgeted bandit problems with continuous random costs". In: *Asian conference on machine learning*, pp. 317–332.

Xia, Yingce, Tao Qin, et al. (2016). "Budgeted multi-armed bandits with multiple plays". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 2210–2216.

Zhang, Jian and Joan Feigenbaum (2006). "Finding Highly Correlated Pairs Efficiently with Powerful Pruning". In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. CIKM '06. Arlington, Virginia, USA: ACM, pp. 152–161. ISBN: 1-59593-433-2. DOI: 10.1145/1183614.1183640. URL: http://doi.acm.org/10.1145/1183614.1183640.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Apprentissage Efficient dans les Problèmes de Semi-Bandits Stochastiques Combinatoires

Mots clés: Bandit combinatoire, régions de confiance, efficience computationnelle

Résumé: Les problèmes de semi-bandits stochastiques combinatoires se présentent naturellement dans de nombreux contextes où le dilemme exploration/exploitation se pose, tels que l'optimisation de contenu web (recommandation/publicité en ligne) ou encore les méthodes de routage à trajet minimal. Ce problème est formulé de la manière suivante : un agent optimise séquentiellement une fonction objectif inconnue et bruitée, définie sur un ensemble puissance $\mathcal{P}([n])$. Pour chaque ensemble A testé, l'agent subit une perte égale à l'écart espéré par rapport à la solution optimale tout en obtenant des observations lui permettant de réduire son incertitude sur les coordonnées de A. Notre objectif est d'étudier l'efficience des politiques pour ce problème, en nous intéressant notamment aux deux aspects suivants : l'efficience statistique, où le critère considéré est le regret subi par la politique (la perte cumulée) qui mesure la performance d'apprentissage ; et l'efficience computationnelle (i.e., de calcul). Il est parfois difficile de réunir ces deux aspects dans une seule politique. Dans cette thèse, nous proposons différentes directions pour améliorer l'efficience statistique, tout en essayant de maintenir l'efficience computationnelle des politiques. Nous avons notamment amélioré les méthodes optimistes en développant des algorithmes d'approximation et en affinant les régions de confiance utilisées. Nous avons également exploré une alternative aux méthodes optimistes, à savoir les méthodes randomisées, et avons constaté qu'elles constituent un candidat sérieux pour pouvoir réunir les deux types d'efficience.

Title: Efficient Learning in Stochastic Combinatorial Semi-Bandits

Keywords: Combinatorial bandit, confidence regions, computational efficiency

Abstract: Combinatorial stochastic semi-bandits appear naturally in many contexts where the exploration/exploitation dilemma arises, such as web content optimization (recommendation/online advertising) or shortest path routing methods. This problem is formulated as follows: an agent sequentially optimizes an unknown and noisy objective function, defined on a power set $\mathcal{P}([n])$. For each set A tried out, the agent suffers a loss equal to the expected deviation from the optimal solution while obtaining observations to reduce its uncertainty on the coordinates from A. Our objective is to study the efficiency of policies for this problem, focusing in particular on the following two aspects: statistical efficiency, where the criterion considered is the regret suffered by the policy (the cumulative loss) that measures learning performance; and computational efficiency. It is sometimes difficult to combine these two aspects in a single policy. In this thesis, we propose different directions for improving statistical efficiency, while trying to maintain the computational efficiency of policies. In particular, we have improved optimistic methods by developing approximation algorithms and refining the confidence regions used. We also explored an alternative to the optimistic methods, namely randomized methods, and found them to be a serious candidate for combining the two types of efficiency.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France