Contributions to a Theory of Pure Exploration in Sequential Statistics

Antoine Barrier

THÈSE

en vue de l'obtention du grade de Docteur, délivré par

l'ÉCOLE NORMALE SUPÉRIEURE DE LYON

École Doctorale N° 512
École Doctorale en Informatique, Mathématiques de Lyon (InfoMaths)

Discipline : Mathématiques

Soutenue publiquement le 20/07/2023, par :

Antoine BARRIER

Contributions à une théorie de l'exploration pure en statistique séquentielle

---

Devant le jury composé de :

MAILLARD, Odalric-Ambrym	Chargé de recherche	INRIA, Univ. de Lille	Rapporteur
VERNADE, Claire	Professeure	Université de Tübingen	Rapporteure
PERCHET, Vianney	Professeur des universités	ENSAE (CREST)	Examinateur
GARIVIER, Aurélien	Professeur des universités	CNRS, ÉNS de Lyon	Directeur de thèse
STOLTZ, Gilles	Directeur de recherche	CNRS, Univ. Paris-Saclay	Co-directeur

Remerciements

Chers Aurélien et Gilles, merci de m'avoir fait confiance pour ces trois années d'aventures ! De par vos différences dans votre approche de la recherche et dans vos habitudes de travail, j'ai appris à vos côtés plus que je ne l'aurais imaginé. Aurélien, je suis toujours impressionné par ta vitesse de raisonnement et par ta maîtrise de tant de domaines des statistiques : une mise en contexte ultra succincte te suffit pour jongler d'un problème à un autre. Tu as su instaurer une ambiance de travail très agréable au sein du petit (mais grossissant) groupe de statisticiens de l'UMPA grâce aux réunions hebdomadaires. Gilles, ton efficacité et ton organisation à toute épreuve m'interrogent sur ton appartenance à l'espèce humaine : 100% de mails répondus dans la demi-heure (peu importe l'heure et le jour d'envoi), une qualité de rédaction inégalable (tout comme ton sens de la formule) et une précision infinie lors de tes relectures (merci pour toutes les typos corrigées dans ce manuscrit et tout au long de cette thèse) . . . Tout simplement épatant et inspirant !

Claire et Odalric, merci à tous les deux d'avoir accepté la lourde tâche de rapporter ma thèse, j'en suis honoré. Vos nombreux commentaires ont contribué à améliorer le manuscrit et à aboutir à cette version finale.

Vianney, je suis également très heureux de te compter parmi mon jury. En fait, tu es celui qui m'a initié à l'apprentissage statistique et à l'intelligence artificielle à Cachan, avec tes exposés en L3 et surtout ton cours en M1. C'est donc une belle manière de boucler la boucle !

Mes remerciements ne sauraient se restreindre à ces trois années de thèse. Avant elles, c'est un long parcours qui a mené à l'élaboration du projet, où tant de personnes sont intervenues. Je profite de ces débuts de remerciements pour mettre en avant les professeurs passionnés et investis ---en mathématiques ou non--- que j'ai eu la chance d'avoir depuis mon enfance.

Les années de prépa à Champo ne me remémorrent que de bons souvenirs, grâce notamment à de formidables enseignants. Parmi eux, il y a bien sûr JML, qui a plus qu'outrepassé ses fonctions depuis plusieurs années : tes nombreux conseils depuis la fin de la prépa m'ont beaucoup aidé, particulièrement durant cette thèse, et j'ai hâte de continuer à discuter de tes projets de sobriétés environnementale et numérique autour d'un bon burger.

Les études cachannaises ont aussi été mathématiquement très fructueuses, un merci particulier à Frédéric Pascal qui a été notre référent et interlocuteur préférentiel pendant ces quatre années, et à Dieter, mon encadrant de stage de M1, pour sa bienveillance et pour m'avoir guidé vers Aurélien.

Cette thèse n'aurait pas été si agréable sans un environnement de travail adéquat. L'ensemble des membres de l'UMPA ainsi que les étudiants du département de mathématiques (avec bien sûr la meilleure des promos d'agrégatifs cette année) y ont amplement contribué.

Je remercie les membres permanents pour toutes les discussions et tous les conseils, notamment Grégory et Thomas avec lesquels j'ai préparé mes enseignements. Une pensée à Claude, exemple de dévouement et de passion, me vient évidemment.

Merci aux secrétaires Jessica, Laure, Magalie et Virginia pour leur sympathie et leur invincible efficacité.

Merci à tous les membres de l'équipe de statistiques d'Aurélien : Alexandre, Aymen, Élise, Hugues, Mehrasa, Pierre et Tomáš. Il est toujours intéressant de discuter de bandits, de RL, ou d'autres problèmes de statistiques avec vous.

Je ne compte plus les moments de vie vécus (au laboratoire ou non) avec les membres non permanents : merci à Alexandre, Aymen, Basile, Benoît, Denis, Cécile, Charlie, Corentin, Héloïse, Hugues, Jules, Juliana,

Léo, Matthieu, Paul, Raphaël (x2), Riccardo, Ronan, Simon, Thomas, Valentine, Vanessa, Vianney, William et tant d'autres … Un merci particulier aux incroyables co-chargés de TD Juliana et Léo, et aux co-locataires du bureau 404 (dont certains cherchent encore l'existence) Hugues[1] *et Basile, qui ont mis une belle ambiance (parfois peu propice au travail …) ces derniers temps, mais aussi avant eux Matthieu et Felipe.*

Au-delà du laboratoire, j'ai eu la chance de participer à de nombreux séminaires qui furent tous de beaux moments de rencontres avec de jeunes chercheurs : je remercie (la liste est loin d'être exhaustive) Annette, Dorian, El Mehdi, Emmanuel, Eugenio, Joseph, Julien, Juliette, Marc, Mete, Quentin et Thibault. Marc, merci également pour nos tentatives de collaboration cette année et pour tous les bons moments passés au CIRM et à Singapour !

Si ces trois années sont passées si vite, c'est grâce à un environnement mathématique stimulant, mais aussi et surtout grâce aux moments passés entre amis ou en famille qui ont rythmé cette thèse et parsemé tant de souvenirs.

Cette thèse fut l'occasion de multiplier les activités sportives. En tête de liste, la découverte d'innombrables terrains de jeux alpins. Mention spéciale aux sorties (((rando-)))trail, vélo et ski de fond avec la team lyonnaise. En tête Rémy avec ses performances XXL à ski de fond, et son taux de nitro imbattable dans les montées de cols. Mais aussi Mattéo avec ses chaussures de trail dernier cri, ses cargaisons de sel et bien sûr sa force légendaire pour soulever des lits. Ou encore Pierre dont la capacité à soigner ses chutes ---aussi bien à vélo que dans les fameuses Pierrades--- n'a pas d'égal (et ça c'est tip top ! Ou idyllique, je ne sais plus …).

Que de beaux souvenirs me reviennent également des treks dans les Écrins avec les champions Fanny, Marie, Paul et Rémi, le Mercantour avec les boss Rémy et Mattéo, et la Vanoise avec les athlètes Jules, Mattéo, Pomme, Romain et Suzanne. Et à tous les autres qui ont partagé de beaux moments en montagne, à pied ou à vélo : Loïc, Florent, Maël, Hugues, Major, Paul, Quentin, ...

La bonne ambiance avec les pongistes, insufflée à l'AS par Vincent puis Fabrice, mais aussi et surtout au GBTT par tout un club, des jeunes aux moins jeunes (il y aurait tant de monde à citer …), comptent aussi parmi les super moments sportifs de ces dernières années.

Il y a bien sûr des amis de longue date qui ont été, encore et toujours, particulièrement présents tout au long de cette thèse. Loïc, Louane et Nathan, vous êtes juste formidables … Il s'en est encore passé des choses : les week-ends de l'ambiance, Saint-Mandrier, Imbours, Europa Park, Center Parcs, LouThan Parc[2]*, et tellement d'autres … J'ai hâte de voir la suite du programme !*

Je rajoute à cette liste les parisiens qui m'ont accueilli lors de mes passages dans la capitale. Les cachannais Anatole le Magicien (en tournée dans tout Paris), Doudouche le hipster, Jobic l'indescriptible et Farf à la bonne humeur inébranlable. Les anciens de Champo ne sont pas en reste : Major et le rituel apéro / pâtes au saumon / film à Bourg-la-Reine, Fanny qui répond toujours présente dès qu'il est possible de se voir (mille mercis pour ta curiosité malgré la distance cette année !), les sessions de jeux chez Ben et Cam's, et Stan et son intrépide aventure dans le monde de la finance. Last but not least, mon cousin Simon, toujours si accueillant avec ses bonnes bouteilles et à l'origine de tant de bons moments lors de mes années parisiennes !

Merci aussi aux infatigables Andy et Benoît, toujours prêts à organiser mille et un traquenards.

J'ai aussi eu la chance d'être proche de ma famille pendant cette thèse. Leur soutien a été constant et très important pour moi. Je remercie mes parents Nathalie et Thierry, mon frère Mathis, mais aussi mes grands-parents lyonnais Claudette et René qui ont tous tant œuvrés pour que je ne manque jamais de rien.

Et enfin, Berthine, mille mercis pour ton soutien ces derniers mois. ¡Espero que éste sea el comienzo de una gran aventura !

---

[1] *oups, je fais partie des personnes qui oublient malheureusement si souvent les « h » de ton nom et prénom, cher Hughues van Hassel …*

[2] *c'est comme ça que je nommerai désormais votre appartement, tant il est synonyme de divertissement et d'amusement !*

Et pour toi, cher lecteur, qui n'ira pas beaucoup plus loin dans la lecture de ce manuscrit, voici un petit pêle-mêle (comme dirait l'autre) de quelques formules anonymes entendues durant ces trois années :

« Un mathématicien c'est pas comme un joueur d'échec, il a le droit de se tromper. »

« Si tu étais à Ninja Warrior, je dirais que tu es au pied du mur. »

« En recherche, on ne fait que jouer au loto. »

« Tes maîtres de thèse sont tes boulets. »

« L'intelligence artificielle, c'est pas des maths. »

Pour terminer, sans doute la question la plus posée ces trois dernières années :

« Qu'est-ce que ça donne pour des bandits Gaussiens à trois bras ? »

Sommaire · Contents

Résumé

Contributions à une théorie de l'exploration pure en statistique séquentielle

Cette thèse, à la croisée entre les domaines de l'intelligence artificielle, de la statistique séquentielle et de l'optimisation, s'intéresse au problème d'identification du meilleur bras (en espérance) dans les bandits non structurés à K bras. Ce problème possède deux approches dont les niveaux de compréhension sont très différents.

Le cadre à confiance fixée est le mieux compris : des stratégies asymptotiquement optimales sont connues, et l'on s'intéresse à l'obtention de garanties non asymptotiques pour des stratégies (si possible) simples et naturelles. Avec des bandits Gaussiens, nous proposons l'analyse à risque fini d'une nouvelle stratégie (asymptotiquement optimale) grâce aux propriétés de régularité de ce modèle. Cette stratégie modifie subtilement la règle d'attribution des tirages de l'algorithme Track-and-Stop en une règle plus prudente et interprétable. Dans le contexte plus général d'un modèle exponentiel, nous proposons l'ébauche d'une analyse de l'asymptotique optimalité d'algorithmes de type Top-Two adaptatifs, dont les règles de choix de tirages sont particulièrement simples.

Par ailleurs, dans le cadre à budget fixé, où l'existence d'une hypothétique complexité reste à démontrer, nous proposons des généralisations à des modèles non-paramétriques des bornes (supérieures et inférieures) connues jusqu'à présent pour des modèles très spécifiques. Les bornes obtenues font intervenir des quantités de théorie de l'information plus précises que les écarts entre les moyennes qui apparaissaient précédemment. Ces nouvelles quantités pourraient être la clé pour mesurer la complexité de l'identification de meilleur bras à budget fixé.

Mots-clés. Problèmes de bandits · Identification de meilleur bras · Statistiques séquentielles · Apprentissage statistique · Intelligence artificielle

Liste des publications

Le présent manuscrit s'appuie sur les deux publications suivantes :

A. Barrier, A. Garivier, and T. Kocák. A Non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022

A. Barrier, A. Garivier, and G. Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 136–181. PMLR, 2023

Les chapitres 3 et 4 reprennent le contenu du papier AIStats 2022, qui porte sur l'étude d'une nouvelle stratégie d'identification de meilleur bras à confiance fixée, `Exploration-Biased-Sampling`, pour des variables gaussiennes. Des guaranties à risque fini sont notamment démontrées, en utilisant de nouveaux résultats de régularité du problème d'optimisation définissant les poids optimaux.

Ensuite, le chapitre 5 présente les travaux issus de la publication ALT 2023. L'objectif est de généraliser les bornes connues en identification de meilleur bras à budget fixé à des modèles généraux, possiblement non paramétriques, en faisant intervenir de nouvelles quantités de théorie de l'information plus précises que les gaps.

Le chapitre 6, qui s'intéresse aux propriétés asymptotiques des algorithmes `top-two` adaptatifs en confiance fixée, est une ébauche contenant des résultats préliminaires non publiés.

Abstract

Contributions to a Theory of Pure Exploration in Sequential Statistics

This thesis lies in the fields of artificial intelligence, sequential statistics and optimization. We focus on the problem of best (in expectation) arm identification in unstructured muti-armed bandits. This problem has two approaches with very different levels of understanding.

The fixed-confidence framework is the best understood: asymptotically optimal strategies are known, and we are interested in obtaining non-asymptotic guarantees for (if possible) simple and natural strategies. Working with Gaussian bandits, we propose a finite risk analysis of a new (asymptotically optimal) strategy using the regularity properties of this model. This strategy slightly modifies the sampling rule of the Track-and-Stop algorithm into a more conservative and interpretable rule. In the more general context of an exponential model, we propose a preliminary analysis of the asymptotic optimality of adaptive Top-Two algorithms, whose sampling rules are particularly simple.

Independently, in the fixed-budget framework, for which the existence of a hypothetical complexity remains to be demonstrated, we propose generalizations to non-parametric models of the existing bounds (upper and lower) that were available so far only for very specific models. The obtained bounds involve more precise information-theoretic quantities than the gaps (differences between the means) which appeared previously. These new quantities could be the key to measuring the complexity of fixed-budget best-arm identification.

Keywords. Multi-Armed Bandits · Best-Arm Identification · Sequential Statistics · Statistical Learning · Machine Learning

Origin of the materials

This manuscript is based on the following two publications:

A. Barrier, A. Garivier, and T. Kocák. A Non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022

A. Barrier, A. Garivier, and G. Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 136–181. PMLR, 2023

Chapters 3 and 4 follow the contents of the AIStats 2022 paper, which deals with the study of a new fixed-confidence best-arm identification strategy, `Exploration-Biased-Sampling`, for Gaussian variables. In particular, finite-risk guarantees are demonstrated, using new regularity results for the optimization problem defining the optimal weights.

Then, Chapter 5 presents the work resulting from the ALT 2023 publication. The aim is to generalize the known bounds in best-arm identification with a fixed budget to general, possibly non-parametric, models, by involving new quantities of information theory that are more precise than the gaps.

Finally, Chapter 6, which deals with the asymptotic properties of adaptive `top-two` algorithms in the fixed-confidence setting, is a draft containing preliminary unpublished results.

CHAPITRE 1

Vue d'ensemble des résultats

Ce chapitre initial résume l'ensemble des contributions de la thèse en les replaçant dans le contexte bibliographique de l'identification de meilleur bras. Après une présentation du problème, nous expliquons comment une profonde compréhension du modèle Gaussien dans le cadre à confiance fixée nous a permis de définir une stratégie avec des garanties non asymptotiques. Puis, dans le cadre à budget fixé, nous énonçons des généralisations de bornes existantes à des modèles quelconques en introduisant de nouvelles mesures de complexité à base de quantités de théorie de l'information.

Sommaire

1.1. Exploration pure dans les problèmes de bandits

Dans cette thèse, on s'intéresse au problème statistique suivant. Un joueur (un *learner*) fait face à $K \geq 2$ distributions de probabilité $\nu_1, \dots, \nu_K$ inconnues. À chaque instant, il peut observer une réalisation de la distribution de son choix, en utilisant les observations passées. Son objectif est d'identifier rapidement la distribution de plus grande espérance. Il s'agit d'un problème d'optimisation séquentielle aux nombreuses applications.

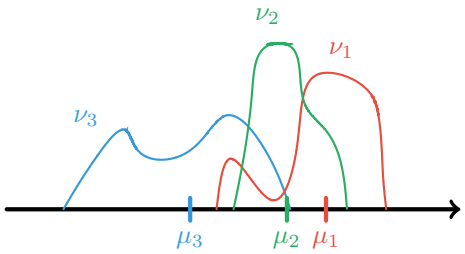


Figure 1.1: μ_a désigne l'espérance de ν_a. Considérant les trois distributions dont nous représentons les fonctions de masse, l'objectif du joueur est d'identifier ν_1 comme étant la meilleure distribution.

Par exemple, en médecine, si l'on souhaite déterminer quel médicament est le plus efficace pour traiter une maladie parmi un ensemble de K médicaments, on peut organiser un essai clinique où l'on administre à chaque malade un des médicaments et l'on observe s'il guérit ou non. On peut alors procéder de manière séquentielle en choisissant le médicament administré à un patient en fonction de toutes les observations précédentes. Si, pour les premiers patients, on voudra observer les médicaments uniformément en l'attente de données fiables, il faudra en cours d'essai clinique se concentrer au fur et à mesure sur les médicaments les plus prometteurs empiriquement. L'efficacité de l'essai clinique dépendra de la stratégie adoptée, et donc de l'arbitrage choisi entre exploration de l'ensemble des distributions et focalisation sur les distributions les plus prometteuses.

Problèmes de bandits. Le cadre et la terminologie utilisés pour modéliser la situation sont ceux des bandits (Lattimore and Szepesvári, 2020) : on se donne un ensemble $\underline{\nu} = (\nu_1, \dots, \nu_K)$ de distributions appartenant à un modèle donné $\mathcal{D}$, où $K \geq 2$ est fixé. Bien que les motivations initiales concernaient des essais cliniques (Thompson, 1933), le terme de *problème de bandit* est utilisé pour désigner $\underline{\nu}$ et provient du nom de machines à sous, les *bandits manchots*, dans les casinos. Les indices des distributions (et par extension leurs distributions associées) sont quant à eux les *bras* du problème $\underline{\nu}$. On désigne par $\underline{\mu} = (\mu_1, \dots, \mu_K)$ le vecteur de moyennes de $\underline{\nu}$, où $\mu_a = \mathrm{E}(\nu_a)$ est l'espérance de ν_a. Dans l'ensemble de la thèse, nous travaillons avec le modèle élémentaire des bandits *non structurés*, i.e., dont les distributions sont supposées indépendantes. Parmi les modèles de distributions possibles, on peut notamment citer :

- le modèle $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ des distributions Gaussiennes avec variance commune $\sigma^2 > 0$, le modèle $\mathcal{D}_{\mathcal{B}}$ des distributions de Bernoulli, ou plus généralement les modèles exponentiels,
- le modèle non paramétrique $\mathcal{P}[0,1]$ des lois à valeurs dans $[0,1]$, ou plus généralement tout modèle $\mathcal{D}_{\sigma^2}$ constitué de variables σ^2–sous-Gaussiennes, i.e., de lois ν satisfaisant la borne suivante sur la fonction des moments, où $X \sim \nu$:

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\Big[e^{\lambda(X-\mathbb{E}[X])}\Big] \leq \mathrm{e}^{\frac{\lambda^2\sigma^2}{2}} \ .$$

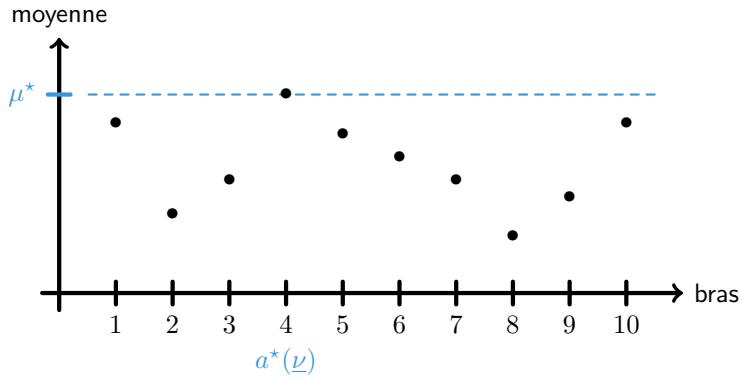


Figure 1.2: Bras optimal et moyenne optimale d'un problème de bandit $\underline{\nu}$.

Identification de meilleur bras. Le problème présenté dans le premier paragraphe est connu sous le nom d'*identification de meilleur bras* ou d'*exploration pure* : l'objectif est d'identifier le bras, supposé unique et noté $a^\star(\underline{\nu})$, dont l'espérance est la plus élevée :

$$\left\{a^\star(\underline{\nu})\right\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{1\leq a\leq K} \mu_a \,.$$

Dans l'ensemble de ce chapitre, nous considérons des problèmes de bandits ayant un unique bras *optimal*. Pour cela, le joueur donne, après un nombre (potentiellement aléatoire) τ d'observations, une estimation $\hat{a}_\tau \in [K] \stackrel{\text{def}}{=} \{1,\dots,K\}$ du meilleur bras $a^\star(\underline{\nu})$. Ce problème peut être traité sous deux approches différentes en fonction des applications.

- Une première possibilité est de travailler à *confiance fixée* (Even-Dar et al., 2006) : le joueur ne peut pas décider de s'arrêter tant qu'il n'est pas certain d'avoir identifié le bras optimal avec un niveau de risque inférieur à un seuil $\delta \in (0,1)$ fixé. On dit alors que sa stratégie est δ–correcte. Le calcul du niveau de risque d'une stratégie dépend des hypothèses sur les distributions, i.e., sur le modèle $\mathcal{D}$ auxquelles elles appartiennent. Ces hypothèses sont connues du joueur. Parmi toutes les stratégies δ-correctes, on voudrait trouver celles qui minimisent le nombre nécessaire d'observations.
- La deuxième approche est donnée par le cadre du *budget fixé* (Audibert et al., 2010) : le joueur est obligé de donner son estimation après un nombre total d'observations T fixé. On cherche les stratégies qui minimisent la probabilité d'erreur, c'est-à-dire la probabilité d'avoir une mauvaise estimation du bras optimal.

De nombreuses applications existent pour les deux approches, selon le nombre d'observations possibles : le budget fixé est adapté pour des essais cliniques où le nombre de patients à disposition est difficilement extensible, alors que la confiance fixée peut être utilisée, par exemple, par des sites web souhaitant tester diverses configurations et disposant de nombreux internautes (test A/B).

De manière assez surprenante, les deux approches conduisent à des connaissances théoriques et empiriques très différentes. Nous allons voir que les stratégies d'apprentissage diffèrent sensiblement.

1.2. Identification de meilleur bras à confiance fixée

Nous étudions d'abord l'identification de meilleur bras à confiance fixée : étant donné $\delta \in (0,1)$ fixé, la stratégie du joueur doit s'arrêter après un temps aléatoire τ_δ de sorte de pouvoir garantir que son estimation $\hat{a}_{\tau_\delta}$ est correcte avec risque δ. L'objectif est de trouver des stratégies δ–correctes qui nécessitent le moins d'observations possibles.

Stratégie. Pour commencer, nous définissons formellement ce qu'est une stratégie. On se donne une suite $(U_t)_{t\geq 0}$ de variables aléatoires i.i.d. de loi $\mathcal{U}([0,1])$, indépendante de tout autre source d'aléa, et on note Y_t l'observation au temps $t \geq 1$. La stratégie du joueur est définie par la donnée :

- d'une *règle d'échantillonnage*, qui décide du choix du bras $A_t \in [K]$ à observer à l'étape $t \geq 1$, en fonction des observations passées $I_{t-1} \overset{\text{def}}{=} (U_0, Y_1, U_1, Y_2, U_2, \dots, Y_{t-1}, U_{t-1})$; A_t est donc $\mathcal{F}_{t-1}$–mesurable, où $\mathcal{F}_{t-1} = \sigma(I_{t-1})$.
- d'une *règle d'arrêt* τ_δ, qui est un temps d'arrêt par rapport à $(\mathcal{F}_t)_{t\geq 0}$,
- d'une *règle de décision* $\hat{a}_{\tau_\delta}$ qui est $\mathcal{F}_{\tau_\delta}$–mesurable et décide quel bras est estimé comme étant optimal (il s'agit le plus souvent du bras de meilleure moyenne empirique).

La stratégie est dite δ–*correcte* si elle vérifie

$$\forall \underline{\nu} \text{ dans } \mathcal{D}, \qquad \mathbb{P}_{\underline{\nu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\nu})) \leq \delta \,,$$

où $\mathbb{P}_{\underline{\nu}}$ désigne la probabilité sous le problème de bandit $\underline{\nu}$.

Notations. Pour un problème de bandit $\underline{\nu}$ fixé, on note $\Delta_a(\underline{\nu}) = \mu^\star - \mu_a$ le *gap* du bras a, c'est-à-dire l'écart entre la moyenne optimale et la moyenne du bras a. Pour une stratégie donnée face au problème $\underline{\nu}$, on note $N_a(t)$ et $\hat{\mu}_a(t)$ le nombre de tirages et la moyenne empirique[1] du bras a à la fin de l'étape t :

$$N_a(t) \overset{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = a\} \qquad \text{and} \qquad \hat{\mu}_a(t) \overset{\text{def}}{=} \frac{1}{N_a(t)} \sum_{s\in[t]} Y_s \, \mathbb{I}\{A_s = a\} \,,$$

où $\mathbb{I}\{E\}$ est l'indicatrice de l'évènement E.

1.2.1. Borne inférieure pour un modèle exponentiel

Un bon critère pour mesurer la performance de ces stratégies est le nombre moyen de tirages $\mathbb{E}_{\underline{\nu}}[\tau_\delta]$. Nous allons voir qu'une notion d'optimalité existe en effet pour cette quantité. Dans un premier temps, Garivier and Kaufmann (2016) ont montré une borne inférieure sur le nombre moyen de tirages d'une stratégie δ–correcte pour un modèle exponentiel. Elle repose sur une inégalité de théorie de l'information faisant intervenir la divergence de Kullback-Leibler.

Remarque. Les premières stratégies d'identification de meilleur bras à confiance fixée, comme la stratégie d'élimination de Even-Dar et al. (2006), garantissent des bornes sur τ_δ avec forte probabilité. Cependant, il n'y a pas de borne inférieure pour ce critère.

La divergence de Kullback-Leibler. La *divergence de Kullback-Leibler* est une pseudo-distance sur les distributions de probabilité qui joue un rôle important en théorie de l'information. Elle est définie, pour toute paire de probabilités $\mathbb{P}$ et $\mathbb{Q}$ définies sur un espace mesurable $(\Omega, \mathcal{A})$, par

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) \overset{\text{def}}{=} \begin{cases} \displaystyle\int_\Omega \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \, \mathrm{d}\mathbb{P} = \mathbb{E}_{X\sim\mathbb{P}}\left[\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)\right] & \text{si } \mathbb{P} \ll \mathbb{Q} \,, \\ +\infty & \text{sinon,} \end{cases}$$

où $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ est la dérivée de Radon-Nikodym de $\mathbb{P}$ par rapport à $\mathbb{Q}$ lorsque $\mathbb{P}$ est absolument continue par rapport à $\mathbb{Q}$. On peut vérifier que la KL est toujours positive mais n'est pas symétrique, et que $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = 0$ si et seulement si $\mathbb{P} = \mathbb{Q}$.

[1] Toutes les stratégies observent chaque bras une fois initialement, donc $\hat{\mu}_a(t)$ est bien défini pour $t \geq K$.

L'inégalité fondamentale. Les bornes inférieures de la littérature sur les bandits utilisent régulièrement des changements de mesure : on considère un problème de bandit $\underline{\zeta}$, dit *alternatif*, dont le bras optimal diffère de celui de $\underline{\nu}$ et on souhaite quantifier combien d'observations sont nécessaires pour discriminer $\underline{\nu}$ et $\underline{\zeta}$. Dans notre cadre, il est pratique d'utiliser l'inégalité de théorie de l'information suivante, qui évite l'utilisation de changements de mesure plus explicites. Une nouvelle preuve de ce résultat, utilisant une martingale arrêtée, est donnée en Section 2.2.2.

Lemme 1.1. [Kaufmann et al., 2016, Lemma 1]
Soient $\underline{\nu}$ et $\underline{\zeta}$ deux problèmes de bandits. Soit une stratégie d'identification de meilleur bras telle que le temps d'arrêt τ est $\mathbb{P}_{\underline{\nu}}$–intégrable. Alors, pour tout évènement E dans $\mathcal{F}_\tau$, on a

$$\sum_{a\in[K]} \mathbb{E}_{\underline{\nu}}[N_a(\tau)]\,\mathrm{KL}(\nu_a,\zeta_a) \geq \mathrm{KL}\Big(\mathrm{Ber}(\mathbb{P}_{\underline{\nu}}(E)),\,\mathrm{Ber}(\mathbb{P}_{\underline{\zeta}}(E))\Big)\,.$$

Modèles exponentiels. On désigne par modèle exponentiel, et on note $\mathcal{D}_{\exp}$, toute famille exponentielle canonique à un paramètre (voir Lehmann and Casella, 1998, Section 1.5). Il s'agit d'un ensemble de distributions ν_θ indexées par $\theta \in \Theta \subset \mathbb{R}$, toutes absolument continues par rapport à une mesure ρ sur $\mathbb{R}$, de densités données par

$$x \longmapsto \frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) \overset{\text{def}}{=} \exp(\theta x - b(\theta))\,,$$

pour une fonction de normalisation b au moins deux fois dérivable. Citons par exemple les distributions de Bernoulli, binomiales, de Poisson, et Gaussiennes de variance commune. On peut montrer que $\mathrm{E}(\nu_\theta) = b'(\theta)$ pour tout $\theta \in \Theta$, et donc que les distributions du modèle sont caractérisées par leur espérance. La divergence de Kullback-Leibler peut aussi être paramétrée par les espérances : on note, pour tout $\theta, \theta' \in \Theta$,

$$d(\mathrm{E}(\nu_\theta), \mathrm{E}(\nu_{\theta'})) \overset{\text{def}}{=} \mathrm{KL}(\nu_\theta, \nu_{\theta'})\,.$$

Cela définit une divergence d qui est strictement convexe et différentiable sur $\mathcal{M}\times\mathcal{M}$, où $\mathcal{M} = (\mu_-, \mu_+)$ est l'intervalle supposé ouvert des moyennes du modèle. En particulier, d est continue, telle que $d(\mu,\mu') = 0$ si et seulement si $\mu = \mu'$, et, pour tout $\mu \in \mathcal{M}$, $d(\mu, \cdot\,)$ et $d(\cdot\,, \mu)$ sont strictement décroissantes sur $(\mu_-, \mu]$, et strictement croissantes sur $[\mu, \mu_+)$.

La divergence de Kullback-Leibler entre variables de Bernoulli jouant un rôle particulier dans les bornes inférieures, on la note kl et on a l'expression suivante :

$$\forall p,q \in (0,1), \qquad \mathrm{kl}(p,q) \overset{\text{def}}{=} \mathrm{KL}\Big(\mathrm{Ber}(p),\mathrm{Ber}(q)\Big) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}\,.$$

Borne inférieure pour un modèle exponentiel. En utilisant le Lemme 1.1 et la propriété de stratégie δ–correcte, Garivier and Kaufmann (2016) ont prouvé la borne inférieure suivante pour un modèle exponentiel, où les problèmes de bandits $\underline{\nu}$ sont caractérisés par leur moyennes $\underline{\mu}$ comme expliqué ci-dessus, et où l'on définit le simplexe

$$\Sigma_K \overset{\text{def}}{=} \Big\{\underline{v} \in [0,1]^K : v_1 + \cdots + v_K = 1\Big\}\,,$$

et l'ensemble des bandits *alternatifs* au problème $\underline{\mu}$:

$$\mathrm{Alt}(\underline{\mu}) \overset{\text{def}}{=} \Big\{\underline{\lambda} \text{ in } \mathcal{D}_{\exp} : a^\star(\underline{\lambda}) \neq a^\star(\underline{\mu})\Big\}\,.$$

Théorème 1.1. [Garivier and Kaufmann, 2016, Theorem 1]
Soit $\mathcal{D}_{\exp}$ un modèle exponentiel. Soit $\delta \in (0, \frac{1}{2})$. Pour toute stratégie δ–correcte, et pour tout problème de bandit $\underline{\mu}$ dans $\mathcal{D}_{\exp}$ avec un unique bras optimal,

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta] \geq T(\underline{\mu}) \operatorname{kl}(\delta, 1-\delta) \geq T(\underline{\mu}) \log \frac{1}{2.4\delta}\,, \tag{1.1}$$

où $T(\underline{\mu})$ est le temps caractéristique *de $\underline{\mu}$, défini par*

$$T(\underline{\mu})^{-1} \stackrel{\text{def}}{=} \sup_{\underline{v} \in \Sigma_K} \inf_{\underline{\lambda} \in \operatorname{Alt}(\underline{\mu})} \sum_{a \in [K]} v_a \, d(\mu_a, \lambda_a)\,. \tag{1.2}$$

Un vecteur de poids optimal. Garivier and Kaufmann (2016) ont montré que le problème d'optimisation (1.2) admet un unique maximiseur $\underline{w}(\underline{\mu})$, appelé le *vecteur de poids optimal* de $\underline{\mu}$:

$$\Big\{\underline{w}(\underline{\mu})\Big\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\underline{v} \in \Sigma_K} \inf_{\underline{\lambda} \in \operatorname{Alt}(\underline{\mu})} \sum_{a \in [K]} v_a \, d(\mu_a, \lambda_a)\,.$$

Pour obtenir une stratégie dont la performance est proche de la borne inférieure (1.1), il faut que les proportions moyennes de tirages $\big(\mathbb{E}_{\underline{\mu}}[N_a(\tau_\delta)]/\mathbb{E}_{\underline{\mu}}[\tau_\delta]\big)_{a \in [K]}$ de cette stratégie soient proches de $\underline{w}(\underline{\mu})$. Autrement dit, si la stratégie ne tire pas les bras selon des fréquences bien précises, elle ne pourra pas être performante.

1.2.2. Résolution du problème d'optimisation définissant $T(\underline{\mu})$

Considérons toujours un modèle exponentiel $\mathcal{D}_{\exp}$. Nous allons voir que la solution du problème d'optimisation (1.2) définissant $T(\underline{\mu})$ est cruciale pour déterminer l'efficacité d'une stratégie.

Coûts de transport. Le problème d'optimisation (1.2) définissant $T(\underline{\mu})$ peut être écrit sous la forme suivante, où $a^\star = a^\star(\underline{\mu})$:

$$T(\underline{\mu})^{-1} = \sup_{\underline{v} \in \Sigma_K} \min_{a \neq a^\star} \underbrace{v_{a^\star} d(\mu^\star, \overline{\mu}_{a^\star, a, \underline{v}}) + v_a d(\mu_a, \overline{\mu}_{a^\star, a, \underline{v}})}_{\stackrel{\text{def}}{=} \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})}\,,$$

$$\text{où} \qquad \overline{\mu}_{a^\star, a, \underline{v}} \stackrel{\text{def}}{=} \frac{v_{a^\star}\mu^\star + v_a \mu_a}{v_{a^\star} + v_a}\,.$$

La quantité $\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})$ s'interprète comme un coût de transport représentant la difficulté de changer les distributions des bras $a^\star$ et a de sorte que a devienne optimal, étant données les fréquences de tirage $\underline{v}$. La preuve de l'unicité de $\underline{w}(\underline{\mu})$ assure que le vecteur de poids optimal égalise les coûts de transport :

$$\forall a \neq a^\star, \qquad \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w}(\underline{\mu})) = T(\underline{\mu})^{-1}\,.$$

Approximation du vecteur de poids optimal. Il est important de pouvoir approcher la valeur du vecteur de poids optimal $\underline{w}(\underline{\mu})$ et de comprendre sa dépendance vis-à-vis du paramètre $\underline{\mu}$. Cela peut notamment permettre de définir des stratégies (voir la présentation de `Track-and-Stop` dans la section suivante). Garivier and Kaufmann (2016) ont montré que le calcul de $\underline{w}(\underline{\mu})$ revenait à déterminer la racine d'une fonction croissante à une variable. On peut donc approcher $\underline{w}(\underline{\mu})$ avec une précision arbitraire en utilisant une méthode de dichotomie. Cette caractérisation de $\underline{w}(\underline{\mu})$ a également permis d'obtenir la continuité de $\underline{\mu} \mapsto \underline{w}(\underline{\mu})$. Cependant, d'autres caractérisations pourraient

permettre d'obtenir des méthodes d'approximation plus efficaces d'un point de vue computationnel et de meilleurs résultats de régularité. Dans cette thèse, nous travaillons sur ce problème d'optimisation dans le cas spécifique d'un modèle Gaussien au Chapitre 3 et, indépendamment, dans un cadre exponentiel général au Chapitre 6.

Caractérisation pour des variables Gaussiennes. Dans le Chapitre 3, on s'intéresse au problème d'optimisation pour des variables aléatoires Gaussiennes standards[2]. Dans ce cas, le problème s'écrit simplement :

$$T(\underline{\mu})^{-1} = \sup_{\underline{v} \in \Sigma_K} \frac{1}{2} \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \Delta_a^2 \,,$$

où l'on note, tant qu'il n'y a pas d'ambiguïté, $\Delta_a = \Delta_a(\underline{\mu})$. Nous montrons que le vecteur de poids optimal $\underline{w}(\underline{\mu})$ est caractérisé par la racine de la fonction $\phi_{\underline{\mu}}$ définie par :

$$\forall r \in \left(\frac{1}{\Delta_{\min}^2}, +\infty \right), \qquad \phi_{\underline{\mu}}(r) \overset{\text{def}}{=} \sum_{a \neq a^\star} \frac{1}{\left(r\Delta_a^2 - 1\right)^2} - 1 \,.$$

On a en effet la proposition suivante.

Proposition 1.2. **[voir Propostion 3.2]**
Soit $\underline{\mu}$ un problème de bandit dont les bras sont Gaussiens standards. Soient $\underline{\Delta} = \underline{\Delta}(\underline{\mu})$, $\underline{w} = \underline{w}(\underline{\mu})$, $T = T(\underline{\mu})$, et soit $r = r(\underline{\mu})$ la solution de $\phi_{\underline{\mu}}(r) = 0$. Alors

$$w_{a^\star} = \frac{1}{1 + \sum_{a \neq a^\star} \frac{1}{r\Delta_a^2 - 1}} \,,$$
$$\forall a \neq a^\star, \qquad w_a = \frac{w_{a^\star}}{r\Delta_a^2 - 1} \,,$$
$$\text{et} \qquad T = 2\frac{r}{w_{a^\star}} \,.$$

Cette caractérisation a plusieurs conséquences pour les modèles Gaussiens à variance commune.

1. Tout d'abord, la fonction $\phi_{\underline{\mu}}$ étant convexe et décroissante, l'application d'une méthode de Newton permet d'approcher r, et donc $\underline{w}(\underline{\mu})$ et $T(\underline{\mu})$, avec une vitesse de convergence quadratique des itérées. Cela accélère la méthode proposée initialement par Garivier and Kaufmann (2016). Voir Section 3.3 pour plus de détails.
2. Nous proposons également de nouvelles bornes pour les valeurs possibles de la coordonnée $w_{a^\star}(\underline{\mu})$ du vecteur de poids optimal et du temps caractéristique $T(\underline{\mu})$:

$$\frac{1}{1 + \sqrt{K-1}} \leq w_{a^\star}(\underline{\mu}) \leq \frac{1}{2} \,,$$
$$\text{et} \qquad \max\left(\frac{8}{\Delta_{\min}^2}, 4\frac{1 + \sqrt{K-1}}{\overline{\Delta^2}} \right) \leq T(\underline{\mu}) \leq 2\frac{\left(1 + \sqrt{K-1}\right)^2}{\Delta_{\min}^2} \,,$$

   où $\overline{\Delta^2} \overset{\text{def}}{=} \frac{1}{K-1} \sum_{a \neq a^\star} \Delta_a^2$ est la moyenne des carrés des gaps et $\Delta_{\min} \overset{\text{def}}{=} \min_{a \neq a^\star} \Delta_a$ est le gap minimal. Toutes ces bornes sont atteintes par des problèmes spécifiques.
3. Nous établissons des résultats de monotonicité lorsque l'on bouge l'un (ou plusieurs) des bras de $\underline{\mu}$. Par exemple, en augmentant la valeur de la moyenne d'un bras sous-optimal a (voir

[2] Ou plus généralement de variance commune $\sigma^2 > 0$.

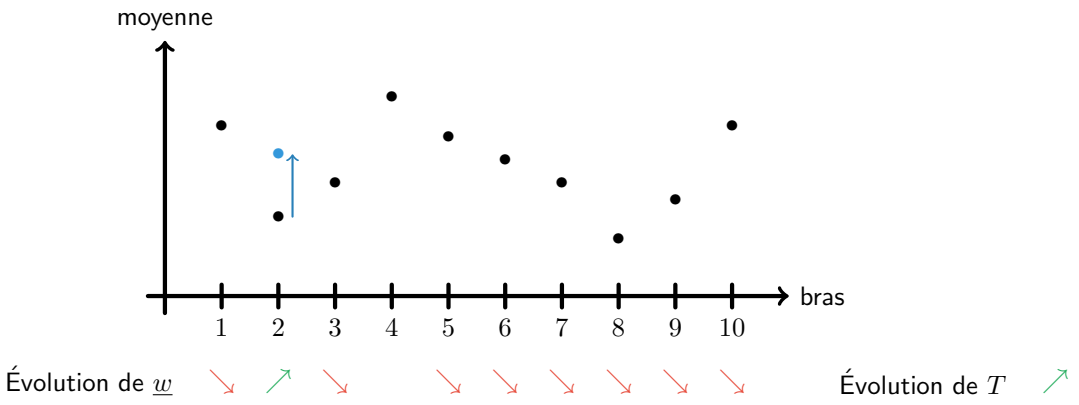


Figure 1.3: En augmentant la moyenne d'un bras sous-optimal, son poids optimal augmente tandis que les poids des autres bras sous-optimaux diminuent, et le temps caractéristique augmente.

Figure 1.3), on augmente son poids $w_a(\underline{\mu})$ tout en diminuant le poids des autres poids sous-optimaux. Comme le gap entre le bras a et le bras optimal diminue, le problème d'identification se complexifie et donc la valeur de $T(\underline{\mu})$ augmente. La Section 3.4 donne l'ensemble des résultats de monotonicité obtenus, qui vont se révéler importants pour la définition de stratégies exploratrices comme la stratégie `Exploration-Biased-Sampling` proposée au Chapitre 4.

4. Enfin, nous démontrons des résultats quantitatifs de régularité des solutions au problème d'optimisation dans la Section 3.5. On montre notamment que les fonctions $\underline{\mu} \mapsto \underline{w}(\underline{\mu})$ et $\underline{\mu} \mapsto T(\underline{\mu})$ sont localement lipschitziennes :

> **Théorème 1.2. [voir Theorem 3.9]**
> *Soient $\underline{\mu}$ et $\underline{\mu}'$ deux problèmes de bandit dont les bras sont Gaussiens standards, et de même bras optimaux. Soient $\underline{\Delta}$, $\underline{w}$ et T (respectivement $\underline{\Delta}'$, $\underline{w}'$ et T') définis comme dans la Proposition 1.2. Supposons que, pour un $\varepsilon \in [0, 1/7]$, on ait*
>
> $$\forall a \neq a^\star, \qquad (1-\varepsilon)\Delta_a^2 \leq {\Delta'_a}^2 \leq (1+\varepsilon)\Delta_a^2 \,.$$
>
> *Alors*
>
> $$\forall a \in [K], \qquad (1-10\varepsilon)w_a \leq w'_a \leq (1+10\varepsilon)w_a \,,$$
> $$\text{et} \qquad (1-3\varepsilon)T \leq T' \leq (1+6\varepsilon)T \,.$$

Cela précise le résultat de continuité de $\underline{\mu} \mapsto \underline{w}(\underline{\mu})$ obtenu[3] par Garivier and Kaufmann (2016). La régularité du problème d'optimisation permet d'envisager de nouvelles techniques de preuves de bornes non asymptotiques, comme celles présentées au Chapitre 4.

Caractérisation pour un modèle exponentiel quelconque. Les techniques développées dans le Chapitre 3 utilisent la forme particulièrement simple de la divergence de Kullback-Leibler entre deux Gaussiennes, et il semble compliqué de les généraliser à tout modèle exponentiel. Dans la Section 6.2, nous donnons cependant une nouvelle caractérisation pour un modèle exponentiel $\mathcal{D}_{\exp}$ quelconque. Celle-ci repose sur l'introduction d'une quantité définie pour $\underline{v} \in \mathrm{int}(\Sigma_K)$, l'intérieur de Σ_K, par

$$T_{\underline{v}} \overset{\text{def}}{=} \sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{a^\star,b,\underline{v}})} \,.$$

[3] Dans le cadre plus général d'un modèle exponentiel.

La caractérisation de $\underline{w}(\underline{\mu})$ faite par Garivier and Kaufmann (2016) assure que $T_{\underline{w}} = T(\underline{\mu})$, et donc, étant donné un vecteur de poids $\underline{v}$, $T_{\underline{v}}$ est une approximation du temps caractéristique. Cela permet de définir une transformation $\underline{W} : \mathrm{int}(\Sigma_K) \to \mathrm{int}(\Sigma_K)$ par

$$\forall \underline{v} \in \mathrm{int}(\Sigma_K), \forall a \in [K], \qquad W_a(\underline{v}) \stackrel{\mathrm{def}}{=} \begin{cases} v_a \frac{T_{\underline{v}}^{-1}}{\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})} & \text{si } a \neq a^\star \,, \\ 1 - \sum_{b \neq a^\star} v_b \frac{T_{\underline{v}}^{-1}}{\mathrm{TC}_{b \to a^\star}(\underline{\mu}, \underline{v})} & \text{si } a = a^\star \,. \end{cases}$$

On peut alors vérifier que $\underline{w}(\underline{\mu})$ est l'unique point fixe de cette transformation.

Proposition 1.3. [voir Proposition 6.2]
Soit $\mathcal{D}_{\mathrm{exp}}$ un modèle exponentiel et $\underline{\mu}$ dans $\mathcal{D}_{\mathrm{exp}}$. Alors $\underline{W}$ a pour unique point fixe $\underline{w}(\underline{\mu})$.

Cette caractérisation a elle aussi des conséquences importantes.

1. Des algorithmes itératifs, utilisant la transformation $\underline{W}$, permettent de faire converger des séquences de vecteurs vers le vecteur de poids optimal. Si les résultats théoriques de convergence restent à établir, nous vérifions que cette convergence a lieu numériquement, quelque soit le problème de bandit et le modèle, avec une convergence assez rapide, qui pourrait là encore accélérer la méthode initiale de Garivier and Kaufmann (2016).
2. Aussi, une interprétation de la transformation $\underline{W}$ mène naturellement à la définition d'un nouvel algorithme de type Top-Two adaptatif, que nous décrirons plus en détails en Section 1.2.5.

1.2.3. L'algorithme `Track-and-Stop`

Pour obtenir des garanties proches de la borne inférieure (1.1), nous avons vu qu'une stratégie doit observer les bras avec des fréquences proches du vecteur de poids optimal (inconnu) $\underline{w}(\underline{\mu})$. L'idée de l'algorithme `Track-and-Stop` de Garivier and Kaufmann (2016) est d'estimer ce vecteur de poids à chaque étape t en utilisant la moyenne empirique disponible $\underline{\hat{\mu}}(t-1)$.

Règle d'échantillonnage. À l'étape t, `Track-and-Stop` choisit donc d'observer le bras qui est le plus en retard par rapport au vecteur de poids $\underline{\hat{w}}(t-1) = \underline{w}(\underline{\hat{\mu}}(t-1))$, on parle de D-tracking, ou à une version cumulative $\frac{1}{t-1}\sum_{s\in[t-1]} \hat{w}_a(s)$ plus prudente, on parle de C-tracking :

$$A_t \in \begin{cases} \mathrm{argmin}_{a\in[K]} N_a(t-1) - \sum_{s\in[t-1]} \hat{w}_a(s) & \text{(C-tracking)}, \\ \mathrm{argmin}_{a\in[K]} N_a(t-1) - (t-1)\hat{w}_a(t-1) & \text{(D-tracking)}. \end{cases}$$

Le biais introduit par l'utilisation du vecteur $\underline{\hat{w}}(t-1)$ plutôt que $\underline{w}(\underline{\mu})$ peut conduire à un sous-échantillonnage de bras de moyennes assez élevées, notamment lorsque les premières estimations sont mauvaises. Pour y remédier, un mécanisme d'exploration forcée vient remplacer la règle d'échantillonnage lorsque l'un des bras n'a pas assez été observé : si le bras le moins observé a été échantillonné moins de $\sqrt{t}$ fois (taux arbiraire), il est tiré automatiquement.

Règle d'arrêt et décision. Le choix des règles d'arrêt est aussi crucial : celles-ci-doivent assurer que la stratégie est δ–correcte sans pour autant nécessiter trop d'observations, ce qui diminuerait la performance de la stratégie. Il semblerait qu'il y ait un consensus autour de la règle d'arrêt du `Global-Likelihood-Ratio` introduite également par Garivier and Kaufmann (2016), qui consiste à stopper l'algorithme dès lors que $Z(t)$ dépasse un certain seuil $\beta(t,\delta)$, où

$$Z(t) \stackrel{\mathrm{def}}{=} \max_{a\in[K]} \min_{b\neq a} Z_{a,b}(t)\,,$$

avec $Z_{a,b}(t)$ le ratio du log–likelihood sous les hypothèses $H_1 : \mu_a > \mu_b$ et $H_0 : \mu_a \leq \mu_b$, étant données les observations à la fin de l'étape t (voir Section 2.2.5 pour plus de détails).

Ils ont notamment établi que cette règle d'arrêt assurait que la stratégie est δ–correcte pour le seuil suivant, quelque soit la règle d'échantillonnage.

Théorème 1.3. [Garivier and Kaufmann, 2016, Proposition 12]
Soit un modèle exponentiel $\mathcal{D}_{\exp}$. Soit $\delta \in (0,1)$ et $\alpha > 1$. Il existe une constante $R = R(\alpha, K)$ telle que, quelque soit la règle d'échantillonnage, la règle d'arrêt du `Global-Likelihood-Ratio` *avec le seuil*

$$\beta(t,\delta) \overset{\text{def}}{=} \log \frac{Rt^{\alpha}}{\delta}\,, \tag{1.3}$$

et avec l'estimation du meilleur bras empirique, assure que la stratégie est δ–correcte.

S'il est encore possible de diminuer légèrement le seuil, il semblerait que l'on ne puisse pas améliorer sensiblement le temps d'arrêt en changeant la condition donnée par le `Global-Likelihood-Ratio`.

Optimalité asymptotique. L'algorithme `Track-and-Stop` est la première stratégie *asymptotiquement optimale*, pour laquelle les garanties lorsque le risque δ tend vers 0 correspondent avec la borne inférieure :

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu})\,.$$

Nous donnons une preuve de ce résultat en Section 6.4 (voir Theorem 6.17).

1.2.4. Des garanties non asymptotiques

L'existence d'une stratégie asymptotiquement optimale questionne sur l'obtention de bornes non asymptotiques. Dans le cadre du modèle $\mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ des distributions Gaussiennes standards avec moyennes dans $[0,1]$, nous définissons au Chapitre 4 une nouvelle stratégie appelée `Exploration-Biased-Sampling`, pour laquelle nous démontrons des bornes à risque fini.

Règle d'échantillonnage. L'idée de cette nouvelle stratégie est de modifier légèrement le fonctionnement de `Track-and-Stop`, en instaurant une région de confiance $\mathcal{CR}_{\underline{\mu}}(t-1)$ autour de $\hat{\underline{\mu}}(t-1)$ afin de définir un vecteur de poids qui favorise et maximise l'exploration, c'est-à-dire qui assure un taux d'observation minimal supérieur à celui donné par le vecteur de poids optimal $\underline{w}(\underline{\mu})$. Pour cela on définit

$$\tilde{\underline{\mu}}(t-1) \in \operatorname*{argmax}_{\underline{\rho} \in \mathcal{CR}_{\underline{\mu}}(t-1)} w_{\min}(\underline{\rho})\,,$$

où $w_{\min}(\underline{\rho}) \overset{\text{def}}{=} \min_{a \in [K]} w_a(\underline{\rho})$ est la composante minimale de $\underline{w}(\underline{\rho})$. Ainsi, dès lors que le problème de bandit $\underline{\mu}$ appartient à la région de confiance $\mathcal{CR}_{\underline{\mu}}(t-1)$, utiliser le vecteur $\tilde{\underline{\mu}}(t-1)$ plutôt que $\underline{w}(\underline{\mu})$ augmente la fréquence minimale d'observation des bras.

La règle d'échantillonnage d'`Exploration-Biased-Sampling` consiste donc à suivre le poids $\tilde{w}(t-1)$ de manière directe (D-tracking) ou cumulative (C-tracking) de la même manière que `Track-and-Stop`. Pour ce faire, il faut justifier que le vecteur $\tilde{w}(t-1)$ est calculable. En utilisant les résultats de monotonicité de la Section 3.4, nous montrons que cela est possible dès lors que la région de confiance est un produit d'intervalles de confiance. Nous choisissons

$$\mathcal{CR}_{\underline{\mu}}(t) \overset{\text{def}}{=} \prod_{a \in [K]} \left[\hat{\mu}_a(t) \pm C_{\frac{\gamma}{K}}(N_a(t))\right], \qquad \text{où} \qquad C_\gamma(s) \overset{\text{def}}{=} 2\sqrt{\frac{\log(\frac{4s}{\gamma})}{s}}\,,$$

où $\gamma \in (0,1)$ est un paramètre de l'algorithme fixé. À mesure que le nombre d'observations augmente, la région de confiance se rétrécit jusqu'à $\{\underline{\mu}\}$ et le biais d'exploration décroît, i.e., $\tilde{\underline{w}}(t) \to \underline{w}(\underline{\mu})$.

Borne non asymptotique. Les garanties théoriques d'`Exploration-Biased-Sampling` sont présentées dans la Section 4.3. Une première remarque intéressante est que cette stratégie a un taux naturel d'exploration : tous les bras sont tirés au moins de l'ordre de $\simeq \sqrt{t}$ fois après t étapes. Cette stratégie ne nécessite donc pas besoin d'exploration forcée (à l'inverse de `Track-and-Stop`). La stratégie est asymptotiquement optimale (comme `Track-and-Stop`), les poids estimés convergeant vers ceux de μ au fur et à mesure que les régions de confiance rétrécissent. Nous démontrons également des garanties non asymptotiques, i.e., à risque fini, lorsque nous conservons la règle d'arrêt `Global-Likelihood-Ratio` de `Track-and-Stop`.

Théorème 1.4. [voir Theorem 4.5]
Soient $\gamma \in (0,1)$, $\alpha \in [1,2]$, $\eta \in (0,1]$, et soit $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$. Il existe un évènement $\mathcal{E}$ de probabilité au moins $1-\gamma$ et $\delta_0 \overset{\text{def}}{=} \delta_0(\underline{\mu}, K, \gamma, \eta, \alpha) > 0$ tels que, pour tout $0 < \delta \leq \delta_0$, l'algorithme `Exploration-Biased-Sampling` *avec le seuil* (1.3) *vérifie*

$$\forall t > (1+\eta)T(\underline{\mu}) \log \frac{1}{\delta}, \qquad \mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) \leq 2Kt \exp\left(-\frac{t w_{\min}(\underline{\mu})}{4T(\underline{\mu})^2} \frac{1}{\log^{\frac{2}{3}} \frac{1}{\delta}}\right),$$

et

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta \mathbb{I}\{\mathcal{E}\}] \leq (1+\eta)T(\underline{\mu}) \log \frac{1}{\delta} + \frac{2^7 K T(\underline{\mu})^4}{w_{\min}(\underline{\mu})^2} \exp\left(-\frac{w_{\min}(\underline{\mu})}{4T(\underline{\mu})} \log^{\frac{1}{3}} \frac{1}{\delta}\right) \log^2 \frac{1}{\delta}.$$

Pour obtenir ces garanties, nous utilisons notamment les résultats de régularité démontrés pour les poids optimaux en Section 3.5, couplés à la stabilité de la procédure de tracking.

Correction des lacunes de `Track-and-Stop`. La stratégie de tracking de `Track-and-Stop` possède quelques limitations. Outre la nécessité de recourir à une exploration forcée, le fort bruitage des observations lors des premiers tours peut rendre les proportions estimées très volatiles, et donc la procédure très aléatoire. De plus, la stratégie ne présente pas le comportement attendu de tirer les bras uniformément tant que trop peu d'informations ont été collectées. Les régions de confiance et la procédure introduite par `Exploration-Biased-Sampling` permettent de corriger ces défauts : l'algorithme commence par une phase d'uniforme exploration (de longueur variable), puis les poids évoluent de manière assez stable. En contrepartie, cela entraîne un fort biais en faveur de l'exploration qui, en pratique, se manifeste par un temps de prise de décision toujours plus long que `Track-and-Stop` (mais d'ordre de grandeur comparable même pour des valeurs modérées de risque).

1.2.5. Les algorithmes Top-Two

De nouvelles stratégies tentent de combler les faiblesses de `Track-and-Stop`, notamment du point de vue du coût computationnel. `Track-and-Stop` nécessite en effet le calcul d'un vecteur de poids optimal à chaque étape. Les algorithmes de type `top-two` introduits par Russo (2016) utilisent des règles d'échantillonnages simples : le bras à tirer est choisi entre un meneur L_t (un *leader*) et un concurrent C_t (un *challenger*). De nombreux choix sont possibles, citons notamment (Jourdan et al., 2022) le meilleur bras empirique pour le leader :

$$L_t \in \operatorname*{argmax}_{a \in [K]} \hat{\mu}_a(t-1),$$

puis, pour le challenger, le bras dont le coût de transport empirique avec le leader est le plus faible :

$$C_t \in \operatorname*{argmin}_{a \neq L_t} \mathrm{TC}_{a \to L_t}\left(\underline{\hat{\mu}}(t-1), \frac{\underline{N}(t-1)}{t-1}\right).$$

Avec un paramètre β non adaptatif. Initialement, ces algorithmes ont été étudiés lorsque le bras à observer est choisi comme étant le leader avec une probabilité $\beta \in (0,1)$ fixe. Comme un bon choix de leader mènera à choisir la plupart du temps le meilleur bras comme leader, cela signifie que ces algorithmes tirent en moyenne le meilleur bras avec proportion β, ce qui ne garantit pas un choix optimal puisque a priori $\beta \neq w_{a^\star}(\underline{\mu})$. Ainsi, les analyses actuelles montrent que certains algorithmes de type Top-Two sont β–asymptotiquement optimaux, car ils vérifient pour des modèles exponentiels :

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T_\beta(\underline{\mu})\,, \qquad \text{où} \qquad T_\beta(\underline{\mu})^{-1} \stackrel{\text{def}}{=} \sup_{\substack{\underline{v}\in\Sigma_K \\ v_{a^\star}=\beta}} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a\, d(\mu_a,\lambda_a)\,,$$

ce qui correspond à la borne inférieure asymptotique pour ces algorithmes. Si l'on peut montrer que des choix raisonnables de β (comme $\beta = \frac{1}{2}$) n'engendrent pas une perte de garantie importante, il n'est pas possible d'obtenir l'optimalité asymptotique des algorithmes Top-Two non adaptatifs.

Avec des proportions adaptatives. Une question naturelle consiste donc à comprendre si l'on peut apprendre, au cours des observations, les proportions idéales de tirages entre le leader et le challenger. La transformation $\underline{W}$ présentée en Section 1.2.2 permet en fait de mettre la main sur de telles proportions, et donc de proposer des algorithmes adaptatifs. Nous montrons en effet que

$$\beta_{L,C} = \frac{w_L\, d(\mu_L, \overline{\mu}_{L,C,\underline{w}})}{\mathrm{TC}_{C\to L}(\underline{\mu},\underline{w})} = \frac{w_L\, d(\mu_L, \overline{\mu}_{L,C,\underline{w}})}{w_L\, d(\mu_L, \overline{\mu}_{L,C,\underline{w}}) + w_C\, d(\mu_C, \overline{\mu}_{L,C,\underline{w}})}\,,$$

est la proportion optimale de tirages du leader L face au challenger C. Cette quantité peut être estimée en remplaçant $\underline{w}$ par les proportions empiriques de tirages $\frac{\underline{N}(t)}{t}$.

Pour un modèle Gaussien avec variances communes, You et al. (2023) ont montré que certains de ces algorithmes Top-Two adaptatifs étaient asymptotiquement optimaux. Cependant, l'analyse reste à établir pour des modèles exponentiels, pour lesquels nous conjecturons l'optimalité asymptotique.

Conjecture 1.4. [voir Conjecture 6.4]
Soit $\mathcal{D}_{\exp}$ un modèle exponentiel. L'algorithme de type `top-two` *adaptatif utilisant le meilleur bras empirique comme leader, puis le challenger minimisant le coût de transport avec le leader, vérifie, pour tout problème de bandit $\underline{\mu}$ dans $\mathcal{D}_{\exp}$ avec moyennes distinctes :*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu})\,.$$

Nous explorons cette conjecture dans la Section 6.3. Les résultats établis dans un cadre Gaussien par You et al. (2023) donnent une feuille de route, et nous expliquons comment adapter leurs arguments dans le cadre général. Deux difficultés majeures sont identifiées : d'une part la propriété d'exploration suffisante (le fait que les bras soient tirés au moins de l'ordre de $\sqrt{t}$ à l'étape t), et d'autre part une relation limite vérifiée par les fréquences empiriques. Si nous conjecturons, en donnant quelques intuitions, que l'exploration suffisante est également vérifiée dans le cadre général (ou peut être forcée si ce n'est pas le cas), la difficulté majeure semble désormais d'obtenir une relation limite similaire au cas Gaussien.

1.3. Identification de meilleur bras à budget fixé

Nous explorons ensuite le cadre à budget fixé, où la stratégie retourne son estimation $\hat{a}_T$ du bras optimal après un nombre déterministe d'observations T. L'objectif est de trouver des stratégies qui minimisent la probabilité d'erreur :

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,.$$

Une décroissance exponentielle. En utilisant l'inégalité de Hoeffding avec des modèles sous-gaussiens, on peut vérifier que la probabilité d'erreur converge vers 0 à vitesse exponentielle pour la stratégie qui tire tous les bras uniformément et retourne le meilleur bras empirique :

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq (K-1)\exp\Big(C(\underline{\nu})T\Big)\,,$$

pour une constante $C(\underline{\nu}) < 0$. On s'intéresse alors à borner le taux de décroissance :

$$\frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,.$$

D'une part, on souhaite obtenir des bornes inférieures, valides pour des suites de stratégies "raisonnables", des quantités

$$\ell(\underline{\nu}) \overset{\text{def}}{=} \liminf_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,.$$

D'autre part, pour une suite de stratégies fixée, on cherche une borne supérieure sur la quantité

$$u(\underline{\nu}) \overset{\text{def}}{=} \limsup_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,.$$

Notons que $\ell(\underline{\nu}) \leq u(\underline{\nu})$ et que ces taux sont négatifs.

Nous allons voir que le cadre à budget fixé est pour l'instant moins bien compris que celui à confiance fixée : les bornes que l'on connaît sur $\ell(\underline{\nu})$ et $u(\underline{\nu})$ ne permettent par d'identifier un taux de décroissance précis, et donc une notion de complexité optimale comme en confiance fixée.

1.3.1. Bornes connues de la littérature

Borne supérieure : l'algorithme `Successive-Rejects`**.** L'une des premières stratégies étudiée à budget fixé est un algorithme d'élimination proposé par Audibert et al. (2010) nommé `Successive-Rejects`. La stratégie consiste à diviser l'exploration en $K-1$ tours. Une liste de bras candidats est initialisée à l'ensemble des bras, et à chaque tour l'ensemble des bras restants sont tirés uniformément, puis le pire bras empirique est éliminé des candidats. Audibert et al. (2010) ont considérées les longueurs de phases, soigneusement choisies et déterministes (non adaptatives), suivantes :

$$\ell_1 \overset{\text{def}}{=} \frac{T}{\overline{\log}\, K}\,, \qquad \text{et} \qquad \forall r \in \{2,\dots,K-1\}, \qquad \ell_r \overset{\text{def}}{=} \frac{T}{(K-r+2)\,\overline{\log}\, K}\,, \tag{1.4}$$

où

$$\overline{\log}\, K \overset{\text{def}}{=} \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k}\,,$$

Ils ont démontré une borne supérieure sur le taux exponentiel limite de décroissance impliquant les gaps. Pour tout problème de bandit $\underline{\nu}$ avec un unique bras optimal dans un modèle σ^2–sous-Gaussien,

$$\limsup_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{4\sigma^2\,\overline{\log}\, K}\min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k}\,, \tag{1.5}$$

où $\overline{\log}\, K$ est de l'ordre de $\log K$, et les gaps sont ordonnés selon $0 = \Delta_{(1)} < \Delta_{(2)} \leq \cdots \leq \Delta_{(K)}$.

Bornes inférieures. La borne supérieure (1.5) dépend crucialement du choix des longueurs de phases, qui a en fait été guidé par le souhait d'obtenir un même terme de complexité que celui obtenu dans une borne inférieure pour un modèle de Bernoulli avec des moyennes dans $[p, 1-p]$. Une preuve technique leur a permis de montrer que, intuitivement, pour tout problème de bandit $\underline{\nu}$ avec un unique bras optimal :

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\frac{5}{p(1-p)} \min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k} \,. \tag{1.6}$$

Le choix des longueurs de phases de `Successive-Rejects` a ainsi permis d'obtenir la même quantité $\min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k}$ dans la borne supérieure, au prix de l'apparition d'un facteur supplémentaire $\overline{\log} K$ dû à la contrainte de budget. Autrement dit, les deux bornes ne correspondent pas.

Si d'autres bornes inférieures ont été proposées dans le cas très spécifique d'un modèle Gaussien (pour lequel la divergence de Kullback-Leibler est symétrique), ou dans le cas de $K = 2$ bras (voir Kaufmann et al., 2016), les bornes inférieures et supérieures connues aujourd'hui ne correspondent pas en toute généralité. Pour résumer, la compréhension du problème à budget fixé est donc bien moindre qu'à confiance fixée :

- les bornes supérieure et inférieure ne correspondent pas, avec un écart multiplicatif de l'ordre de $\log K$,
- les bornes existantes sont formulées uniquement à partir des gaps entre les distributions (et non pas de quantités plus précises comme la divergence de Kullback-Leibler),
- les bornes (inférieures) sont valables pour des modèles très spécifiques (Gaussiens et Bernoulli).

1.3.2. Généralisations des bornes existantes à des modèles quelconques

Si le premier point de la liste ci-dessus semble délicat à traiter (voir Section 1.3.3), les techniques de preuves connues peuvent être généralisées à de nombreux modèles (incluant par exemple des modèles exponentiels et non paramétriques), comme nous le montrons au Chapitre 5. Les généralisations de ces bornes font intervenir de nouvelles quantités de théorie de l'information, plus informatives que les gaps, et l'obtention des bornes inférieures se fait grâce à des hypothèses naturelles sur le comportement attendu d'une bonne stratégie.

De nouvelles quantités de théorie de l'information. La borne inférieure de théorie d'information donnée par le Lemme 1.1 peut également être utilisée dans le cadre à budget fixé. Cependant, pour obtenir des bornes sur le taux exponentiel de décroissance, les rôles de $\underline{\nu}$ et de l'alternative $\underline{\zeta}$ doivent être inversés, i.e., on utilise l'inégalité

$$\sum_{a\in[K]} \mathbb{E}_{\underline{\zeta}}[N_a(T)] \,\mathrm{KL}(\zeta_a, \nu_a) \geq \mathrm{kl}\Big(\mathbb{P}_{\underline{\nu}}(E),\, \mathbb{P}_{\underline{\zeta}}(E)\Big) \,.$$

pour tout évènement $E \in \mathcal{F}_T$. Si en confiance fixée, le temps caractéristique s'exprimait en fonction des $\big(\mathrm{KL}(\nu_a, \zeta_a)\big)_{a\in[K]}$, comme par exemple dans (1.2), il semblerait en conséquence que les bornes sur les taux de décroissance à budget fixé s'expriment plutôt en fonction des $\big(\mathrm{KL}(\zeta_a, \nu_a)\big)_{a\in[K]}$, autrement dit les arguments des divergences de Kullback-Leibler sont inversés.

Pour un modèle général $\mathcal{D}$, on introduit alors, pour $\nu \in \mathcal{D}$ et $x \in \mathbb{R}$, les quantités

$$\mathcal{L}_{\inf}^{<}(x,\nu) = \inf\big\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \;\text{ s.t. }\; \mathrm{E}(\zeta) < x\big\} \,,$$
$$\text{et} \qquad \mathcal{L}_{\inf}^{>}(x,\nu) = \inf\big\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \;\text{ s.t. }\; \mathrm{E}(\zeta) > x\big\} \,.$$

Nous allons montrer que ces quantités interviennent dans des généralisations des bornes existantes en remplacement des gaps. Elles pourraient ainsi être la clé pour mesurer la complexité du cadre à budget fixé.

Remarque 1.5. Une application de l'inégalité de Pinsker permet de montrer, lorsque les distributions du modèle sont à valeurs dans $[0,1]$, que

$$\forall x \leq \mathrm{E}(\nu), \quad \mathcal{L}_{\inf}^{<}(x,\nu) \geq 2\big(x-\mathrm{E}(\nu)\big)^2, \qquad \text{et} \qquad \forall x \geq \mathrm{E}(\nu), \quad \mathcal{L}_{\inf}^{>}(x,\nu) \geq 2\big(x-\mathrm{E}(\nu)\big)^2.$$

Dans le cas d'un modèle exponentiel $\mathcal{D}_{\exp}$, ces quantités s'écrivent simplement sous la forme de la divergence de Kullback-Leibler d du modèle :

$$\forall x \leq \mathrm{E}(\nu), \quad \mathcal{L}_{\inf}^{<}(x,\nu) = d(x,\mathrm{E}(\nu)), \qquad \text{et} \qquad \forall x \geq \mathrm{E}(\nu), \quad \mathcal{L}_{\inf}^{>}(x,\nu) = d(x,\mathrm{E}(\nu)).$$

Une analyse plus précise de `Successive-Rejects`. Nous proposons une nouvelle analyse de `Successive-Rejects` qui consiste à remplacer l'application de l'inégalité de Hoeffding dans l'analyse de Audibert et al. (2010) par une simple application de la borne de Cramér-Chernoff. Il en résulte le remplacement des gaps dans la borne (1.5) par une quantité dépendant des transformées de Fenchel-Legendre $(\phi^\star_{\nu_a})_{a\in[K]}$ des log–fonctions génératrices des moments des distributions de $\underline{\nu}$.

Remarque. On définit, pour une distribution ν,

$$\forall \lambda \in \mathbb{R}, \quad \phi_\nu(\lambda) = \log \int_{\mathbb{R}} \mathrm{e}^{\lambda x}\, \mathrm{d}\nu(x) \qquad \text{et} \qquad \forall x \in \mathbb{R}, \quad \phi^\star_\nu(x) = \sup_{\lambda\in\mathbb{R}} \{\lambda x - \phi_\nu(\lambda)\},$$

et l'on rappelle que $\phi^\star_\nu$ intervient dans le contrôle des déviations d'un N-échantillon de loi ν par la borne de Cramér-Chernoff :

$$\begin{aligned} &\forall x \leq \mathrm{E}(\nu), \qquad \mathbb{P}(\overline{X}_N \leq x) \leq \exp\big(-N\,\phi^\star_\nu(x)\big), \\ \text{et} \quad &\forall x \geq \mathrm{E}(\nu), \qquad \mathbb{P}(\overline{X}_N \geq x) \leq \exp\big(-N\,\phi^\star_\nu(x)\big). \end{aligned}$$

On obtient le résultat suivant, où l'on définit, pour tout $\nu, \nu' \in \mathcal{D}$ avec $\mathrm{E}(\nu') < \mathrm{E}(\nu)$,

$$\Phi(\nu',\nu) \overset{\text{def}}{=} \inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]} \{\phi^\star_{\nu'}(x) + \phi^\star_\nu(x)\}.$$

Proposition 1.6. **[voir Corollary 5.4]**
Soit un modèle $\mathcal{D}$. La séquence de stratégies `Successive-Rejects` *utilisant les longueurs de phase* (1.4) *vérifie, pour tout problème de bandit $\underline{\nu}$ dans $\mathcal{D}$ avec un unique bras optimal,*

$$\limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{\overline{\log}\, K} \min_{2\leq k\leq K} \frac{\Phi(\nu_{\sigma_k},\nu^\star)}{k}. \tag{1.7}$$

où l'on range les bras a par ordre croissant de $\Phi(\nu_a,\nu^\star)$, i.e., on considère la permutation σ telle que

$$0 = \Phi(\nu_{\sigma_1},\nu^\star) < \Phi(\nu_{\sigma_2},\nu^\star) \leq \ldots \leq \Phi(\nu_{\sigma_{K-1}},\nu^\star) \leq \Phi(\nu_{\sigma_K},\nu^\star).$$

Sous des conditions de régularité assez légères du modèle $\mathcal{D}$ considéré (voir Section 5.4.2), nous montrons que la quantité $\Phi(\nu',\nu)$ est égale à la quantité

$$\mathcal{L}(\nu',\nu) \overset{\text{def}}{=} \inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]} \Big\{\mathcal{L}_{\inf}^{>}(x,\nu') + \mathcal{L}_{\inf}^{<}(x,\nu)\Big\}, \tag{1.8}$$

ce qui permet de réécrire la borne supérieure (1.7) en fonction des quantités $\mathcal{L}_{\inf}^{<}$ et $\mathcal{L}_{\inf}^{>}$:

$$\forall \underline{\nu} \text{ dans } \mathcal{D}, \qquad \limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{\overline{\log}\, K} \min_{2\leq k\leq K} \frac{\mathcal{L}(\nu_{\sigma_k},\nu^\star)}{k}. \tag{1.9}$$

Pour un modèle sous-Gaussien, cela améliore la borne supérieure (1.5) puisque l'on peut prouver, via l'inégalité de Pinsker, que $\mathcal{L}(\nu',\nu) \geq \frac{1}{4\sigma^2}\big(\mathrm{E}(\nu)-\mathrm{E}(\nu')\big)^2$.

De nouvelles bornes inférieures. Nous montrons également plusieurs bornes inférieures, qui dépendent de différentes hypothèses sur les suites de stratégies considérées. Une hypothèse générale est la *consistence*, c'est-à-dire le fait que l'identification a lieu asymptotiquement presque sûrement quelque soit le problème de bandit, i.e., que $\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \underset{T\to+\infty}{\to} 0$ pour tout $\underline{\nu}$ dans $\mathcal{D}$.

Une première borne est obtenue en faisant les hypothèses supplémentaires suivantes :

- *équilibrage pour le pire bras* : le bras de plus petite espérance est tiré avec fréquence au plus $\frac{1}{K}$ asymptotiquement,
- *exploitation astucieuse de l'élagage des bras sous-optimaux* : la suite de stratégies, doublement indexée par K et T, est plus efficace pour identifier le meilleur bras si l'on enlève un des bras sous-optimaux du problème.

Sous ces deux hypothèses, nous montrons la borne suivante.

Théorème 1.5. [voir Theorem 5.10]
Soit un modèle $\mathcal{D}$. Soit une suite de stratégies, doublement indexée, qui est consistante, équilibrée pour le pire bras, et qui exploite astucieusement l'élagage des bras sous-optimaux. Alors, pour tout problème de bandit $\underline{\nu}$ dans $\mathcal{D}$ avec des moyennes distinctes,

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \frac{\mathcal{L}^{<}_{\inf}(\mu_{(k)}, \nu^\star)}{k}\,, \tag{1.10}$$

où les bras sont ordonnés selon $\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$.

Ce résultat généralise la borne inférieure (1.6) de Audibert et al. (2010) à tout modèle $\mathcal{D}$. Dans le cas particulier d'un modèle de Bernoulli, l'inégalité de Pinsker permet en effet (cf. Remarque 1.5) de déduire (1.6) de la borne (1.10).

Cette borne inférieure ne s'exprime pas en termes d'infima de combinaisons de $\mathcal{L}^{>}_{\inf}$ et $\mathcal{L}^{<}_{\inf}$, i.e., sous la forme $\mathcal{L}$ introduite en (1.8) et que l'on retrouve dans la borne supérieure de `Successive-Rejects` (1.9). Sans parvenir à obtenir une borne inférieure impliquant la quantité $\mathcal{L}$, nous proposons tout de même une borne inférieure pour des modèles dits *normaux*[4] et des suites de stratégies dites *monotones*, qui limitent (asymptotiquement) les fréquences de tirages des bras sous-optimaux :

$$\forall \underline{\nu} \text{ dans } \mathcal{D}, \forall a \in [K], \qquad \limsup_{T\to+\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_{(a)}(T)]}{T} \leq \frac{1}{a}\,.$$

Théorème 1.6. [Voir Theorem 5.13]
Considérons un modèle normal $\mathcal{D}$. Soit une séquence de stratégies consistante et monotone sur $\mathcal{D}$. Alors, pour tout problème de bandit $\underline{\nu}$ dans $\mathcal{D}$ avec des moyennes distinctes,

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \min_{2\leq j\leq k} \inf_{x\in[\mu_{(j)},\mu_{(j-1)})} \left\{ \frac{\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})}{j-1} + \frac{\mathcal{L}^{<}_{\inf}(x, \nu^\star)}{j} \right\}.$$

Cette borne implique notamment la borne plus lisible suivante

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \inf_{x\in[\mu_{(k)},\mu_{(k-1)})} \left\{ \frac{\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})}{k-1} + \frac{\mathcal{L}^{<}_{\inf}(x, \nu^\star)}{k} \right\}, \tag{1.11}$$

[4] Qui nécessitent uniquement que l'infimum définissant $\mathcal{L}^{>}_{\inf}(x, \nu)$ soit atteint par une suite de distributions dont les espérances tendent vers x.

et, de surcroît, la borne inférieure (1.10) en prenant $x = \mu_{(k)}$. La borne (1.11) va dans la direction du terme de complexité qui apparaît dans la borne supérieure (1.9) avec la quantité $\mathcal{L}$, cependant l'infima est pris sur de plus petits intervalles.

Remarque. Les bornes présentées au Chapitre 5 n'améliorent pas l'écart entre les bornes inférieure et supérieure : il y a toujours (au moins) une constante multiplicative de l'ordre de $\overline{\log} K \simeq \log K$ entre les deux.

1.3.3. Existence d'une complexité

La différence de facteur de l'ordre $\log K$ entre les bornes inférieure (1.6) et supérieure (1.5) amène à se demander s'il existe, comme en confiance fixée, une notion de complexité dans le cadre à budget fixé. Si des résultats minimax comme la borne inférieure de Carpentier and Locatelli (2016) montrent que les bornes inférieures ne sont peut-être pas encore optimales, on se demande également s'il existe une stratégie *uniformément* optimale, qui puisse atteindre une borne inférieure asymptotique simultanément sur l'ensemble des problèmes de bandits.

Des travaux récents tentent de démontrer qu'une telle stratégie n'existe pas pour des modèles assez vastes. Degenne (2023) a notamment démontré que, pour un modèle Gaussien, il n'existe pas de stratégie qui soit, uniformément sur chaque problème de bandit, aussi performante que la meilleure stratégie tirant les bras avec des proportions fixes contre ce problème de bandit. Ce comportement, très différent du cadre à confiance fixée, pourrait s'expliquer par la difficulté que peuvent avoir les stratégies à se fier aux observations dans ce cadre.

Notation

This symbol in the external margin highlights an important notation or convention for the remaining of the thesis.

All tuples (vectors of $\mathbb{R}^K$, list of distributions, ...) are underlined.

Remark. For the sake of clarity, some technical notations of Chapters 3 to 5 are not aggregated here. For the same reasons, most of the pieces of notation of Chapter 6 are not included, as the chapter presents work-in-progress ideas with abundant notation.

Generalities

This first table contains general notation. Each piece of notation is associated with a reference to the page where it is defined or where it appears first.

Notation	Designation	Page
K	Positive integer (number of arms)	32
$\underline{v}$	Vector of $\mathbb{R}^K$	44
$[K]$	Set $\{1, \dots, K\}$	33
$\mathrm{int}(A)$	Interior of a set A	48
sgn	Sign function	50
Σ_K	Simplex of dimension $K-1$	44
$\mathfrak{S}_K$	Set of permutations of $[K]$	62
$\overline{\log} K$	$\frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k}$	61

Probabilities

$X \sim \nu$	The random variable X has law ν	33
$\mathrm{E}(\nu)$	Expectation of a distribution ν	33
$\mathbb{E}[X]$	Expectation of a random variable X	33
$\mathbb{I}\{E\}$	Indicator function of an event E	35
$\mathbb{P}^X$	Push-forward measure of a random variable X under probability $\mathbb{P}$	40
$\mathbb{P} \ll \mathbb{Q}$	$\mathbb{Q}$ absolutely dominates $\mathbb{P}$	42
$\mathrm{Leb}_{[0,1]}$	Lebesgue measure restricted to $[0,1]$	41
$\mathrm{Supp}(\nu)$	Support of distribution ν	145
$m(\nu)$	$\inf \mathrm{Supp}(\nu)$	145
$M(\nu)$	$\sup \mathrm{Supp}(\nu)$	145
δ_x	Dirac mass at x	146

Bandit problems

Letters a, b, c always refer to arms, that is elements of $[K]$. In subindices for sums and infima, or with quantifiers, we sometimes omit to explicitly mention $[K]$ for simplicity: for example, given a fixed arm b, $\sum_{a \neq b}$ denotes the sum over arms $a \in [K] \setminus \{b\}$.

Notation	Designation	Page
K	Number of arms	32
ν_a	Probability distribution of arm a	32
μ_a	Mean of probability distribution of arm a	33
$\underline{\nu}$	Bandit problem, set of K distributions $(\nu_1, \dots, \nu_K)$	33
$\underline{\mu}$	Mean vector of bandit problem $\underline{\nu}$	44
$a^\star(\underline{\nu})$ $a^\star(\underline{\mu})$	Set of best arms of $\underline{\nu}$, or unique best arm of $\underline{\nu}$	33 44
$\nu^\star$	Optimal distribution of $\underline{\nu}$	33
$\mu^\star$	Optimal mean of $\underline{\nu}$	33
$\mathbb{P}_{\underline{\nu}}$ $\mathbb{P}_{\underline{\mu}}$	Probability under bandit problem $\underline{\nu}$	35 44
$\mathbb{E}_{\underline{\nu}}$ $\mathbb{E}_{\underline{\mu}}$	Expectation under bandit problem $\underline{\nu}$	35 44
Δ_a, $\Delta_a(\underline{\nu})$	Gap of arm a in $\underline{\nu}$	36
$\Delta_{\min}$	Minimal positive gap of $\underline{\nu}$	39
$\Delta_{\max}$	Maximal gap of $\underline{\nu}$	74
$\overline{\Delta^2}$	Average squared gap of $\underline{\nu}$	48
	Alternative	
$\underline{\zeta}$	Another bandit problem (often an alternative to $\underline{\nu}$)	40
$\underline{\lambda}$	Mean of $\underline{\zeta}$	44
$\mathrm{Alt}(\underline{\nu})$ $\mathrm{Alt}(\underline{\mu})$	Set of alternative bandits problems to $\underline{\nu}$	66 44
	Optimization problem	
$T(\underline{\mu})$	Characteristic time of $\underline{\mu}$	44
$T_\beta(\underline{\mu})$	Characteristic time of $\underline{\mu}$ under the constraint that the best arm is pulled a fraction β of times	57
$\underline{w}$, $\underline{w}(\underline{\mu})$	Optimal weight vector of $\underline{\mu}$	47
$w_{\min}$, $w_{\min}(\underline{\mu})$	Minimal component of the optimal weight vector of $\underline{\mu}$	74
$\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})$	Transportation cost of arm a to $a^\star$ in $\underline{\nu}$, given draw proportions $\underline{v}$	46
$\overline{\mu}_{a^\star, a, \underline{v}}$	Weighted mean of $\mu^\star$ and μ_a with weights $v_{a^\star}$ and v_a	46
(a)	a–th arm of $\underline{\nu}$ in terms of means	37
$\underline{\nu}^\sigma$	Permuted bandit instance associated to $\underline{\nu}$ and permutation σ	57
$H(\underline{\nu})$	Complexity function	69
$H_\Sigma(\underline{\nu})$	$\sum_{a \neq a^\star(\nu)} \frac{1}{\Delta_a^2}$	64

Strategies

We use a special font when referring to algorithms and procedures, like for instance `Track-and-Stop` or `sampling-rule`. Algorithm names start with a capital letter, while general procedures do not.

Notation	Designation	Page
δ	Confidence parameter	33
T	Budget parameter	34
A_t	Observed arm at time step t	35
Y_t	Observation at time step t	35
U_t	External randomization at time step t	35
I_t	Available history at time step t	35
$\mathcal{F}_t$	σ-algebra generated by I_{t+1}	35
$N_a(t)$	Number of observations of arm a at time step t	35
$\hat{\mu}_a(t)$	Empirical mean of arm a at time step t	35
$X_{a,n}$	n–th observation of arm a	37
$\hat{\mu}_{a,n}$	Mean of the n first observations of arm a	37
	Stopping and decision rules	
τ	Stopping time of a δ–correct strategy	35
τ_δ		35
$\hat{a}_\tau$, $\hat{a}_{\tau_\delta}, \hat{a}_T$	Estimated best arm	35
$Z_{a,b}(t)$	Generalized $\log$–likelihood ratio	50
$Z(t)$	Test statistic for the `Global-Likelihood-Ratio` stopping rule	50
$\beta(t,\delta)$	Threshold function	50
R	Constant ensuring δ–correctness of threshold (2.27)	51
α	Constant in threshold (2.27)	51
	Elimination strategies	
r	Round	36
S_{r-1}	Set of candidates arms remaining at the beginning of round r	36
a_r	Arm removed at the end of round r	36
n_r	Number of pulls of the arm eliminated at the end of phase r	38
ℓ_r	Length of round r	60
	`Track-and-Stop`	
$\underline{\hat{w}}(t)$	Plug-in estimate of $\underline{w}(\mu)$ at time step t	49
$U(t)$	Set of under-sampled arms at time step t	49
	`Exploration-Biased-Sampling`	
$\mathcal{CR}_{\underline{\mu}}(t)$	Confidence region for $\underline{\mu}$ after t observations	101
$C_\gamma(s)$	Constant defining the width of $\mathcal{CR}_{\underline{\mu}}(t)$	101
$\underline{\tilde{\mu}}(t)$	Biased exploring bandit inside $\mathcal{CR}_{\underline{\mu}}(t)$	102
$\underline{\tilde{w}}(t)$	Optimal weight of $\underline{\tilde{\mu}}(t)$	102
	`top-two` algorithms	
L_t	Leader at time step t	55
C_t	Challenger at time step t	55
β	Probability parameter of choosing the leader in a `top-two` non-adaptive algorithm	57

Information-theoretic quantities

Notation	Designation	Page
$\mathrm{KL}(\mathbb{P},\mathbb{Q})$	Kullback-Leibler divergence between probability measures $\mathbb{P}$ and $\mathbb{Q}$	40
d	Mean-parametrized Kullback-Leibler divergence of an exponential family	43
kl	Mean-parametrized Kullback-Leibler divergence of the Bernoulli family	44
$\mathcal{L}^{<}_{\inf}(x,\nu)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)<x\}$	67
$\mathcal{L}^{\leq}_{\inf}(x,\nu)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)\leq x\}$	126
$\mathcal{L}^{>}_{\inf}(x,\nu)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)>x\}$	67
$\mathcal{L}^{\geq}_{\inf}(x,\nu)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)\geq x\}$	126
$\mathcal{L}(\nu,\nu')$	$\inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]}\left\{\mathcal{L}^{\geq}_{\inf}(x,\nu')+\mathcal{L}^{\leq}_{\inf}(x,\nu)\right\}$	67
$\mathcal{K}^{<}_{\inf}(\nu,x)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)<x\}$	46
$\mathcal{K}^{>}_{\inf}(\nu,x)$	$\inf\{\mathrm{KL}(\zeta,\nu) : \zeta\in\mathcal{D} \text{ s.t. } \mathrm{E}(\zeta)>x\}$	46

Models of distributions

Notation	Designation	Page
$\mathcal{D}$	General set of distributions all admitting a finite mean	33
$\mathcal{P}[0,1]$	Distributions taking values in $[0,1]$	33
$\mathcal{D}_{\mathcal{B}}$	Bernoulli distributions	33
$\mathcal{B}_{[p,1-p]}$	Bernoulli distributions with means in $[p,1-p]$	64
$\mathcal{D}_{\mathcal{N}_{\sigma^2}}$	Gaussian distributions with known variance $\sigma^2>0$	33
$\mathcal{D}^{\mathcal{M}}_{\mathcal{N}_{\sigma^2}}$	Distributions of $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ with means in $\mathcal{M}$	96
$\mathcal{D}^{[0,1]}_{\mathcal{N}_1}$	Standard Gaussian distributions with mean in $[0,1]$	89
$\mathcal{D}_{\sigma^2}$	σ^2–sub-Gaussian distributions	33
	Exponential models	
$\mathcal{D}_{\exp}$	One-canonical exponential family	33
Θ	Natural space parameter	43
θ	Parameter	43
b	Normalizing function	43
$\mathcal{M}=(\mu_-,\mu_+)$	Open interval of the expectations of distributions	43
d	Mean-parametrized Kullback-Leibler divergence	43
kl	Mean-parametrized Kullback-Leibler divergence of the Bernoulli family	44
$C_{\mathcal{D}}$	Regularity constant associated to a model $\mathcal{D}$	65

CHAPTER 2

Pure Exploration in Multi-Armed Bandits

In this preliminary chapter, we first introduce and motivate the statistical problems considered in this thesis, which constitute a sub-domain of the multi-armed bandit literature known as best-arm identification. Then we review the state-of-the-art theoretical knowledge around those problems and highlight our contributions (to be all presented in detail in Chapters 3 to 6).

In the fixed-confidence setting, we explain that the problem is well-understood in the asymptotic regime and we present new results for a Gaussian model which allow us to define a new strategy with non-asymptotic guarantees. We also present partial results for the asymptotic analysis of adaptive `top-two` algorithms.

Then, we highlight the difficulties faced by the literature to obtain a precise comprehension of the fixed-budget setting. In that setting, we present generalizations of existing bounds to possibly non-parametric models. Those bounds are based on new information-theoretic quantities.

Contents

2.1. A Sequential Learning Problem

In this thesis, we study the following sequential learning task. A learner faces a set of $K \geq 2$ unknown real probability distributions $\nu_1, \dots, \nu_K$. At each time step, the learner observes an independent realization of one of the distributions, that she actively chooses according to previous observations. Her objective is to quickly gather evidence so as to identify the distribution of highest mean.
This statistical task is a sequential and active optimization problem, which are sub-domains of machine learning.

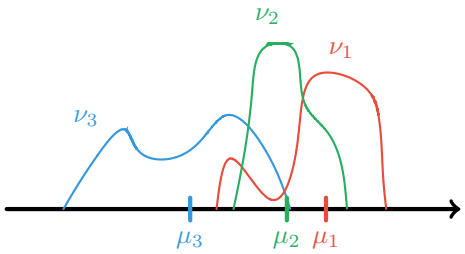


Figure 2.1: μ_a denotes the expectation of ν_a. For the three distributions for which we represent the probability mass functions, the goal of the learner is to identify ν_1 as the best distribution.

Motivations. Before going any further, we give two concrete applications of this learning task.

Application 2.1. [Medical trial]
Assume that you get K candidate treatments to cure a disease. You organize a medical trial to find the treatment that responds best to the disease[1]. You can allocate one treatment to each participant of the trial and observe how she responds to this treatment. You may proceed sequentially so as to choose the next treatment to allocate based on past observations.

Application 2.2. [Online advertising, A/B testing]
You own a webpage that you want to monetize by including an advertisement in it. Your income will be proportional to the number of clicks on the ad. Given K possible ads, you want to find the ad which maximizes the commitment of users. To do so, for each new viewer, you can choose an ad to display, based on previously collected data, and check whether or not he clicks on it.
A famous variant of this problem is A/B testing: you want to find which one among two versions of a webpage maximizes customer satisfaction.

First intuitions on the learner's strategy. The learner's strategy has an impact on the collected data: she chooses the next distribution to observe with the knowledge of previous observations. This is why we say that the learner is *active* and we introduced the problem as a *sequential* learning task. It is in contrast with more "classical" statistical problems for which we cannot influence the data collected. In our problem, how the learner relies on past observations might evolve over time, as intuition suggests:

- during the first steps, the learner has access to a limited amount of observations of each distribution, hence she cannot estimate precisely the associated means. Consequently, it sounds natural to mistrust those observations and proceed to a uniform exploration of the distributions (i.e., observe all of them equally often),
- as more and more observations are gathered, the accuracy of the estimations increases, and the learner should allocate more observations to empirically promising distributions in order to better differentiate them.

[1] In a meaning that needs to be defined: a good treatment should cure the disease while avoiding side effects.

This highlights a major difficulty when dealing with those identification problems: how much can the learner trust previous observations? Does she need to focus on under-sampled distributions (for which she has poor estimates) or on promising distributions (which behave well empirically)?

Multi-armed bandits. The problem introduced in this section is known in the literature as the problem of *best-arm identification* in *multi-armed bandits*. Although the first motivations for those questions concerned medical applications (see Thompson, 1933), this vocabulary comes from casinos and the slot machines sometimes called "one-armed bandits". Considering a set of K machines, a player might try to identify the machine with the best average payoff.

The problem instance $\underline{\nu} = (\nu_1, \dots, \nu_K)$ is called a *K–armed bandit instance* and the indices of its distributions, which belong to $[K] \stackrel{\text{def}}{=} \{1, \dots, K\}$ are its *arms*[2] (hence at each round we choose an arm to observe). All distributions of $\underline{\nu}$ belong to some model denoted by $\mathcal{D}$, which is a subset of the set of real distributions which admit a finite mean.

Example. We may cite, for example,

- the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian distributions with common variance $\sigma^2 > 0$,
- the model $\mathcal{D}_{\mathcal{B}}$ of Bernoulli distributions with means in $(0, 1)$,
- exponential models, denoted by $\mathcal{D}_{\exp}$ (see details in page 43),
- the model $\mathcal{D}_{\sigma^2}$ of σ^2–sub-Gaussian distributions, that is the set of distributions ν such that:

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\Big[e^{\lambda(X - \mathbb{E}[X])}\Big] \leq \mathrm{e}^{\frac{\lambda^2 \sigma^2}{2}}, \tag{2.1}$$

  where $X \sim \nu$,
- the model $\mathcal{P}[0, 1]$ of bounded distributions taking values in $[0, 1]$.

Multi-armed bandit problems have been vastly studied in the literature. The interest in this model comes from the wide variety of applications (see Lattimore and Szepesvári, 2020, Chapter 1 for additional motivations), which has led to the study of many objectives (including regret minimization and best-arm identification), the consideration of different utility functions (mean, quantile, CVaR, etc.), of various structures when there is an underlying dependency between the means (linear bandits, combinatorial bandits, graph bandits, etc.), and also extensions to an infinite number of arms.

Best-arm identification. In the rest of this chapter, unless otherwise specified, we only consider bandit instances that have a unique *best arm* (or *optimal arm*) denoted by $a^\star(\underline{\nu})$:

$$\Big\{a^\star(\underline{\nu})\Big\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{a \in [K]} \mu_a \,,$$

where $\mu_a \stackrel{\text{def}}{=} \mathrm{E}(\nu_a)$ is the mean of arm a. The corresponding distribution and its mean will often be denoted by $\nu^\star$ and $\mu^\star$:

$$\nu^\star \stackrel{\text{def}}{=} \nu_{a^\star(\underline{\nu})} \qquad \text{and} \qquad \mu^\star \stackrel{\text{def}}{=} \mu_{a^\star(\underline{\nu})} \,,$$

and other arms are said to be *sub-optimals*.
The best-arm identification task consists in identifying $a^\star(\underline{\nu})$. As we will detail in the next sections, the problem can be considered under two different approaches:

- in the *fixed-confidence* setting, we introduce some confidence parameter $\delta \in (0, 1)$ and we want to design strategies that will stop after some (random) finite number of observations and give an estimate of the best arm $a^\star(\underline{\nu})$ which is correct with probability at least $1 - \delta$. Such

[2] As the indices characterize the distribution, a slight abuse consists in using arms to directly denote the distributions.

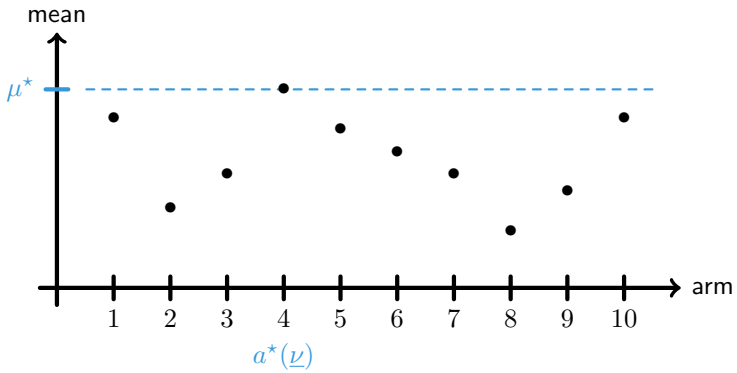


Figure 2.2: Optimal arm and mean of a bandit instance $\underline{\nu}$.

strategies are called δ–correct. Of course, by accessing a very large number of observations, we will be able to design δ–correct strategies, yet we want to avoid unnecessary observations. This is why, in that setting, we try to find δ–correct strategies that minimize the expected number of queries. This setting was originally considered by Even-Dar et al. (2006).

- in the *fixed-budget* setting, we have to give an estimate of the best arm after a given total of T observations. The objective is to design strategies that minimize the probability of misidentification. This setting is studied in Audibert et al. (2010).

Surprisingly, the two settings are not equivalent: they do not lead to the same strategies, and they are not equally understood in the literature (as we will see in this introductive chapter).

Example. Obtaining a δ–correct estimate of the best arm might require a long number of queries, which is not adapted to every situation:

- during a medical trial (Application 2.1), the panel of participants is limited, hence you have a maximal number of possible observations and might model the situation as a fixed-budget best-arm identification problem,
- in online advertising (Application 2.2), if your webpage is visited frequently enough, you have access to a potentially large number of internet users, so that you can wait until obtaining a δ–correct estimate of the best add, and then choose this add to be displayed to all future users.

The problem of best-arm identification is part of the *pure explorationpure exploration* problems (Lattimore and Szepesvári, 2020, Chapter 33): observing sub-optimal distributions is not directly penalized (contrarily to the problem of regret minimization) and even required to increase the confidence on the prediction. However, exploration might be done carefully:

- in the fixed-confidence setting, observing sub-optimal distributions more than necessary will be costly for the total number of queries that we want to minimize,
- in the fixed-budget setting, too many observations of the worst arms will imply that less budget will be devoted to comparing the best distributions, hence increasing the probability of error.

Before being seen as bandit problems, both settings of the best-arm identification problem have been considered in the "*ranking and selection*" literature (see Hong et al., 2021 for a survey).

2.2. Best-Arm Identification with a Fixed-Confidence

In this section, we describe the fixed-confidence setting introduced by Even-Dar et al. (2006) and the main results obtained in the literature for that setting. The objective is to design strategies that after some random number of steps τ return an estimate $\hat{a}_\tau \in [K]$, which is equal to the best arm $a^\star(\underline{\nu})$ with probability at least $1-\delta$, where $\delta \in (0,1)$ is a fixed confidence parameter.

We begin with a mathematical definition of what a strategy is in that setting. We denote by Y_t the observation at step $t \geq 1$ and consider an independent sequence $(U_t)_{t\geq 0}$ of i.i.d. random variables of law $\mathcal{U}([0,1])$ to be used as an external randomization by the strategy.

Formal definition of a strategy. In the fixed-confidence setting, a strategy is defined by

- A *sampling rule*, which consists in choosing the arm $A_t \in [K]$ to observe at each time step $t \geq 1$. This arm A_t depends on the previous observations $Y_1, \dots, Y_{t-1}$, but also possibly on some external randomization that we capture by the random variable U_{t-1}. A_t is thus $\mathcal{F}_{t-1}$–measurable, where

$$\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(I_{t-1})\,, \qquad \text{with } I_{t-1} \stackrel{\text{def}}{=} (U_0, Y_1, U_1, Y_2, U_2, \dots, Y_{t-1}, U_{t-1})\,.$$

  I_{t-1} corresponds to the information available at the end of time step $t-1$, i.e., to the history.
- A *stopping rule* τ, which is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$.
- A *decision rule* $\hat{a}_\tau$ which is $\mathcal{F}_\tau$–measurable.

The general structure of a strategy is presented in Algorithm 1. The stopping rule τ highly depends on the confidence parameter δ, this is why we will denote it as τ_δ from now on. For the sampling and decision rules, most of the presented strategies in this chapter are independent[3] from δ.

For a given strategy facing a bandit problem $\underline{\nu}$, let $N_a(t)$ and $\hat{\mu}_a(t)$ denote the number of pulls and the empirical mean[4] of arm a at step t:

$$N_a(t) \stackrel{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = a\} \qquad \text{and} \qquad \hat{\mu}_a(t) \stackrel{\text{def}}{=} \frac{1}{N_a(t)} \sum_{s\in[t]} Y_s\, \mathbb{I}\{A_s = a\}\,,$$

where $\mathbb{I}\{E\}$ denotes the indicator (or characteristic) function of event E.

A most natural decision rule, used by many strategies, is to return the arm with the best (i.e., the largest) empirical estimate $\hat{\mu}_a(\tau_\delta)$, see Algorithm 2 .

For the sake of clarity, probabilities and expectations will be indexed by the bandit problem under which observations are done: $\mathbb{P}_{\underline{\nu}}$ (respectively, $\mathbb{E}_{\underline{\nu}}$) denotes the probability (respectively, expectation) under the bandit problem $\underline{\nu}$. We now give a concrete definition of δ–correctness.

Definition 2.3. [δ–correct strategy]
Let $\delta \in (0,1)$. A strategy is δ–correct on a model $\mathcal{D}$ if for all K–armed bandit problems $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal arm,

$$\mathbb{P}_{\underline{\nu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\nu})) \leq \delta\,. \tag{2.2}$$

Remark. The condition (2.2) means that the strategy returns a bad arm with controlled probability. As the strategy may never stop, it is not equivalent to ask that the strategy returns $a^\star$ with high probability:

$$\mathbb{P}_{\underline{\nu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} = a^\star(\underline{\nu})) \geq 1-\delta\,. \tag{2.3}$$

Note that both conditions (2.2) and (2.3) are used in the literature, which is not confusing as we will see that efficient strategies satisfy $\tau_\delta < +\infty$ $\mathbb{P}_{\underline{\nu}}$–a.s..

[3] All but elimination algorithms presented in the next section.
[4] All considered strategies initially pull each arm once, hence $\hat{\mu}_a(t)$ is well-defined for $t \geq K$.

Algorithm 1: General structure of a fixed-confidence strategy

Input: confidence parameter δ
`sampling-rule`, `stopping-condition`, `decision-rule`
Output: stopping time τ_δ
estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once // initialization
$t \leftarrow K$
while `stopping-condition`(I_t, δ) *is not satisfied* **do**
  Increase t by 1
  $A_t \leftarrow$ `sampling-rule`(I_{t-1})
  Observe $Y_t \sim \nu_{A_t}$
$\tau_\delta \leftarrow t$
$\hat{a}_{\tau_\delta} \leftarrow$ `decision-rule`(I_{τ_δ})

Algorithm 2: `Empirical-Best` decision rule

Input: history of observations I_{τ_δ}
Output: estimated best arm $\hat{a}_{\tau_\delta}$

Choose $\hat{a}_{\tau_\delta} \in \underset{a \in [K]}{\operatorname{argmax}}\, \hat{\mu}_a(\tau_\delta)$

We will see that the δ–correct condition is ensured by a careful choice of the stopping rule, while the sampling rule determines the performance of the strategy. We explore possible choices of both procedures in Sections 2.2.1, 2.2.5 and 2.2.7.

2.2.1. Naive Approaches

In this section, we present the first δ–correct strategies introduced in the literature by Even-Dar et al. (2006). They considered *elimination* (or *racing*) algorithms, which are strategies with a particular structure: the exploration is split in rounds (or phases). The strategy maintains a list of candidate arms, starting with all arms, and drops (at least) one arm at the end of each phase. We will only consider elimination strategies for which, in each phase r, the random set S_{r-1} of active candidates is uniformly explored, and the worst empirical arm a_r is removed (hence strategies have $K-1$ rounds). See Algorithm 3 for details.

Remark. Elimination algorithms are the only strategies considered in this thesis that do not use the `Empirical-Best` decision rule (Algorithm 2): they recommend the remaining arm of S_{K-1}, which might have a smaller empirical estimate than one of the arms eliminated in rounds 1 to $K-2$ (even if this is quite unlikely).

We consider here the model $\mathcal{P}[0,1]$ of bounded distributions taking values in $[0,1]$, but the results generalize to any σ^2–sub-Gaussian model $\mathcal{D}_{\sigma^2}$. Even-Dar et al. (2006) proposed a δ–correct elimination algorithm for which they obtained guarantees on the number of pulls depending on the *gaps* of bandit problem $\underline{\nu}$. The gap $\Delta_a(\underline{\nu})$ of arm a is the expected difference between the rewards under the optimal distribution $\nu^\star$ and under ν_a:

$$\Delta_a(\underline{\nu}) \stackrel{\text{def}}{=} \mu^\star - \mu_a \,.$$

Algorithm 3: General structure of an elimination algorithm

Input: confidence parameter δ
`stopping-condition` for each phase
Output: stopping time τ_δ
estimated best arm $\hat{a}_{\tau_\delta}$

$t \leftarrow 0$
$S_0 \leftarrow [K]$
for *each phase* $r \in [K-1]$ **do**
  while `stopping-condition`(r, I_t) *does not hold* **do**
    Observe each arm of S_{r-1} once
    Increase t by the cardinality of S_{r-1}
  Choose $a_r \in \operatorname*{argmin}_{a \in S_{r-1}} \hat{\mu}_a(t)$
  $S_r \leftarrow S_{r-1} \setminus \{a_r\}$
$\tau_\delta \leftarrow t$
Define $\hat{a}_{\tau_\delta}$ as the unique element of S_{K-1}

In the sequel, when there is no confusion, we set $a^\star = a^\star(\underline{\nu})$ and $\Delta_a = \Delta_a(\underline{\nu})$ for all $a \in [K]$. We will also order arms by their means, using (reverse) notation of order statistics:

$$\mu_{(1)} > \mu_{(2)} \geq \mu_{(3)} \geq \cdots \geq \mu_{(K)} \,.$$

As a consequence, note that $a^\star(\underline{\nu}) = (1)$.
Letters a, b, c always refer to arms, that is, to elements of $[K]$. In subindices for sums and infima, or with quantifiers, we sometimes omit to explicitly mention $[K]$ for simplicity: for example, given a fixed arm b, $\sum_{a \neq b}$ denotes the sum over arms $a \in [K] \setminus \{b\}$.

To be efficient in the identification task, the racing Algorithm 3 should not eliminate the best arm at the end of all phases. If remaining arms have been pulled n times at the end of a phase, such elimination might occur if two n–sample averages of distributions $\nu^\star$ and ν_a, for some sub-optimal arm a, are in reverse order compared to the expectations of the underlying distributions. We will first focus on the probability of those undesired events, which will then allow us to derive phase lengths that ensure the δ–correctness of the elimination Algorithm 3. Before doing so, we introduce convenient notation offered by optional skipping.

Optional skipping. Let $(X_{a,n})_{a \in [K], n \geq 1}$ be independent random variables such that $X_{a,n} \sim \nu_a$ for all $a \in [K]$ and $n \geq 1$. *Optional skipping* (see Garivier et al., 2022, Section 4.1 for an extensive presentation) is a useful tool in multi-armed bandit problems based on Doob's optional skipping (Doob, 1953, Theorem 5.2), which allows assuming that, given $A_t = a$, the observation at time step t is $Y_t = X_{a,N_a(t)}$. As a consequence, we notably get

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{n=1}^{N_a(t)} X_{a,n} \overset{\text{def}}{=} \hat{\mu}_{a,N_a(t)} \,,$$

making a link between $\hat{\mu}_a(t)$ which is a quantity defined by the global time step t to some $\hat{\mu}_{a,n}$ which depends on local steps of arm a. It might be highlighted that the sequence $(X_{a,n})_{a \in [K], n \geq 1}$ is assumed to be drawn regardless of the strategy. In the analysis, it is sometimes useful to rely on the existence of $\hat{\mu}_{a,n}$ even if arm a has been totally pulled less than n times. See, e.g., Equation (2.7).

Discriminate between two arms. Assume that you have to discriminate the means of arms $a^\star$ and a with confidence δ. How many observations do you need if you pull both arms equally? One can prove that at most $n = \Delta_a^{-2} \log \frac{1}{\delta}$ pulls of each arm are required using Hoeffding's inequality.

Remark. We recall that Hoeffding's inequality states that, if N is an integer and $X_1, \dots, X_N$ are independent random variables such that

- either each X_s belongs to $[a_s, b_s]$ almost surely, with $\sigma_s = \frac{b_s - a_s}{2}$,
- or, more generally[5], each X_s is σ_s^2–sub-Gaussian, see (2.1),

then

$$\forall t > 0, \qquad \mathbb{P}\left(\sum_{s=1}^N X_s - \mathbb{E}[X_s] \geq t \right) \leq \exp\left(- \frac{t^2}{2 \sum_{s=1}^N \sigma_s^2} \right). \tag{2.4}$$

For the choice of n given above, the misidentification probability is indeed controlled by δ:

$$\begin{aligned} \mathbb{P}_{\underline{\nu}}\Big(\hat{\mu}_{a,n} \geq \hat{\mu}_{a^\star,n} \Big) &= \mathbb{P}_{\underline{\nu}}\Big(n(\hat{\mu}_{a,n} - \hat{\mu}_{a^\star,n} + \Delta_a) \geq n\Delta_a \Big) \\ &\leq \exp\Big(-\Delta_a^2 n \Big) \\ &\leq \delta \,, \end{aligned} \tag{2.5}$$

by applying Hoeffding's inequality (2.4) to

$$n(\hat{\mu}_{a,n} - \hat{\mu}_{a^\star,n} + \Delta_a) = \sum_{s=1}^n (X_{a,s} - \mathbb{E}[X_{a,s}]) - \sum_{s=1}^n (X_{a^\star,s} - \mathbb{E}[X_{a^\star,s}]) \,,$$

which is the centered sum of a $2n$–sample of $\frac{1}{4}$–sub-Gaussian variables (as each $X_{a,s}$ and $X_{a^\star,s}$ belongs to $[0,1]$ almost surely). Using this simple observation will allow us to design δ–correct elimination algorithms.

With known gaps. Assume that we know all values of the gaps, but not the mapping $a \mapsto \Delta_a$. Then we can use inequality (2.5) to fix deterministic phase lengths so as to control the probability of misidentification by δ. Indeed, let n_r be the determined and pre-defined total number of pulls at the end of round r of an arm that was still running in that phase (i.e., belonging to S_{r-1}). The algorithm fails to identify the best arm if $a^\star$ is eliminated at some phase, hence the probability of misidentification can be decomposed as follows:

$$\begin{aligned} \mathbb{P}_{\underline{\nu}}(\hat{a}_{\tau_\delta} \neq a^\star) &= \sum_{r \in [K-1]} \mathbb{P}_{\underline{\nu}}(a^\star \text{ is eliminated at round } r) \\ &= \sum_{r \in [K-1]} \mathbb{P}_{\underline{\nu}}(a_r = a^\star) \,. \end{aligned} \tag{2.6}$$

Fixing $r \in [K-1]$, we will choose n_r such that $\mathbb{P}_{\underline{\nu}}(a_r = a^\star) \leq \frac{\delta}{K}$, which will make the strategy δ–correct. We observe that S_{r-1} contains $K - r + 1$ arms, hence at least one arm among the r arms $\{(K), (K-1), \dots, (K-r+1)\}$. We deduce the following bound, using optional skipping and, for

[5] Hoeddfing's lemma exactly claims that, if a random variable belongs almost surely to an interval of length 2σ, then it is σ^2–sub-Gaussian.

the third inequality, bound (2.5):

$$\begin{aligned}
\mathbb{P}_{\underline{\nu}}(a_r = a^\star) &= \mathbb{P}_{\underline{\nu}}\Big(a^\star \in S_{r-1} \text{ and } \hat{\mu}_{a^\star,n_r} \leq \min_{a \in S_{r-1}\setminus\{a^\star\}} \hat{\mu}_{a,n_r}\Big) \\
&\leq \mathbb{P}_{\underline{\nu}}\Big(\exists a \in \{(K),(K-1),\dots,(K-r+1)\} : \hat{\mu}_{a^\star,n_r} \leq \hat{\mu}_{a,n_r}\Big) \qquad (2.7) \\
&\leq \sum_{a=K-r+1}^{K} \mathbb{P}_{\underline{\nu}}\Big(\hat{\mu}_{a^\star,n_r} \leq \hat{\mu}_{(a),n_r}\Big) \\
&\leq \sum_{a=K-r+1}^{K} \exp\Big(-\Delta_{(a)}^2 n_r\Big) \\
&\leq K \exp\Big(-\Delta_{(K-r+1)}^2 n_r\Big) \qquad (2.8) \\
&\leq \frac{\delta}{K}
\end{aligned}$$

as soon as $n_r \geq \left\lceil \frac{2}{\Delta_{(K-r+1)}^2} \log \frac{K}{\delta} \right\rceil$.

Using this lower bound to define the phase lengths we obtain a δ–correct elimination algorithm (relying on the knowledge of the gaps) for which the number of queries is deterministic: as the arm eliminated in phase r is pulled n_r times and the surviving arm is pulled n_{K-1} times, we get

$$\tau_\delta = \sum_{r\in[K-1]} n_r + n_{K-1} = \sum_{a\neq a^\star} \left\lceil \frac{2\log\frac{K}{\delta}}{\Delta_a^2} \right\rceil + \left\lceil \frac{2\log\frac{K}{\delta}}{\Delta_{\min}^2} \right\rceil , \qquad (2.9)$$

where $\Delta_{\min} \overset{\text{def}}{=} \min_{a\neq a^\star} \Delta_a = \Delta_{(2)}$. By observing that

$$2\log\frac{K}{\delta} \sum_{a\neq a^\star} \frac{1}{\Delta_a^2} \leq \tau_\delta \leq 8\log\frac{K}{\delta} \sum_{a\neq a^\star} \frac{1}{\Delta_a^2} ,$$

the sample complexity is of order

$$\mathcal{O}\left(\log\frac{K}{\delta} \sum_{a\neq a^\star} \frac{1}{\Delta_a^2}\right).$$

Remark. As we observed, this strategy has a constant stopping time τ_δ. Elimination algorithms have been similarly studied in fixed-budget best-arm identification (Audibert et al., 2010; Karnin et al., 2013), as we will discuss in Section 2.3.

Without the knowledge of gaps. Without the knowledge of gaps, an elimination strategy has no choice but to consider adaptive phase lengths in order to be δ–correct. Even-Dar et al. (2006) developed such an algorithm, using Algorithm 3 with a stopping criterion for each phase depending on δ, K and t and based on anytime confidence intervals (we will discuss more this concept in the paragraph on stopping conditions in page 50). They proved that their strategy satisfies

$$\tau_\delta = \mathcal{O}\left(\sum_{a\neq a^\star} \frac{\log\frac{K}{\delta\Delta_a}}{\Delta_a^2}\right) = \mathcal{O}\left(\log\left(\frac{K}{\delta\Delta_{\min}}\right) \sum_{a\neq a^\star} \frac{1}{\Delta_a^2}\right) \qquad (2.10)$$

with probability at least $1-\delta$.

2.2.2. The Fundamental Inequality

When measuring the performance of δ–correct strategies, one may be interested in deriving bounds on the number of pulls either with high probability, as in bound (2.10), or in expectation. From now on, we focus on bounds obtained in expectation, as Sections 2.2.3 and 2.2.4 will prove that this criterion seems more suitable.

Garivier and Kaufmann (2016) proved that the performance in expectation of δ–correct strategies cannot be arbitrarily good: they obtained an instance-dependent lower bound on the expected number of pulls of any δ–correct strategy that we will present in Section 2.2.3. The result is based on a fundamental inequality whose applications go beyond the fixed-confidence setting. This fundamental inequality relies on change-of-measure arguments and on information-theoretic properties of the Kullback-Leibler divergence that we briefly introduce in the next paragraph.

The Kullback-Leibler divergence. The *Kullback-Leibler divergence* is a pseudo-distance on probability distributions that appears to play a special role in information theory. It is defined, for all pairs of probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable space $(\Omega, \mathcal{A})$, by

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \displaystyle\int_\Omega \log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\,\mathrm{d}\mathbb{P} = \mathbb{E}_{X\sim\mathbb{P}}\left[\log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)\right] & \text{if } \mathbb{P}\ll\mathbb{Q}\,, \\ +\infty & \text{otherwise}\,, \end{cases}$$

where $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ denotes the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$ when $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. One can easily check that $\mathrm{KL}(\mathbb{P},\mathbb{Q})$ is a non-negative and non-symmetric quantity, and that it is null if and only if $\mathbb{P}=\mathbb{Q}$.
The Kullback-Leibler divergence satisfies a useful inequality to derive lower bounds, namely the *data-processing inequality* (or *contraction of entropy*), which states that the divergence between push-forward measures is smaller than the divergence between the original measures. More precisely, if $\mathbb{P},\mathbb{Q}$ are two probability distributions on a measurable space $(\Omega,\mathcal{A})$ and $X:(\Omega,\mathcal{A})\to(\Omega',\mathcal{A}')$ is a random variable, then

$$\mathrm{KL}\left(\mathbb{P}^X,\mathbb{Q}^X\right) \le \mathrm{KL}(\mathbb{P},\mathbb{Q})\,, \tag{2.11}$$

where $\mathbb{P}^X$ (respectively, $\mathbb{Q}^X$) denotes the push-forward measure of X under $\mathbb{P}$ (respectively, $\mathbb{Q}$). See Ali and Silvey (1966) for proof of this statement, and, e.g., Cover and Thomas (2006) for additional information on the Kullback-Leibler divergence.

The fundamental inequality. Instance-dependent lower bounds can be obtained by performing changes of measure, an argument extensively used in the bandit literature (see, e.g., Lai and Robbins, 1985; Mannor and Tsitsiklis, 2004). This will allow us to quantify how many steps are necessary to discriminate $\underline{\nu}$ from a "close" alternative bandit problem $\underline{\zeta}$ which has a different optimal arm. Kaufmann et al. (2016) derived the following inequality for best-arm identification problems, which is convenient to use as it "hides" the change of measure within a high-level informational argument.

Lemma 2.4. [Kaufmann et al., 2016, Lemma 1]
Let $\underline{\nu}$ and $\underline{\zeta}$ be K–armed bandit problems. Consider any best-arm identification strategy and assume that the stopping-time τ is $\mathbb{P}_{\underline{\nu}}$–integrable. Then, for all events E in $\mathcal{F}_\tau$, we have

$$\sum_{a\in[K]} \mathbb{E}_{\underline{\nu}}[N_a(\tau)]\,\mathrm{KL}(\nu_a,\zeta_a) \ge \mathrm{KL}\Big(\mathrm{Ber}(\mathbb{P}_{\underline{\nu}}(E)),\,\mathrm{Ber}(\mathbb{P}_{\underline{\zeta}}(E))\Big)\,. \tag{2.12}$$

Remark. As we will see, this inequality can be applied both in the fixed-confidence and the fixed-budget settings, but it will have stronger implications in the first one. We will explain in Remark 2.18 that this is mainly due to the fact that, in Equation (2.12), expectations on the number of pulls are relative to bandit problem $\underline{\nu}$; which implies that $\underline{\nu}$ and $\underline{\zeta}$ have very different roles in this inequality (in addition to the fact that the Kullback-Leibler divergence is non-symmetric).

The proof of Kaufmann et al. (2016) is based on explicit changes of measure and Wald's lemma, while Garivier et al. (2019) provided a shorter proof using information-theoretic properties of the Kullback-Leibler divergence when τ is a deterministic time. We will unify both proof schemes by applying a martingale stopping theorem.

Stochastic transition kernels. In the following proof, we will use the fact that the law of I_t under[6] $\mathbb{P}_{\underline{\nu}}$ can be defined iteratively by *regular stochastic transition kernels*, from which we now give a definition. Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be two measurable spaces. We say that $K : \Omega \times \mathcal{A}' \to [0,1]$ is a *regular transition kernel* from $(\Omega, \mathcal{A})$ to $(\Omega', \mathcal{A}')$ if $K(\omega, \cdot)$ is a probability distribution on $(\Omega', \mathcal{A}')$ and if $K(\cdot, B)$ is $\mathcal{A}$–measurable for all $B \in \mathcal{A}'$. For a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$, we define the probability distribution $K\mathbb{P}$ on $(\Omega \times \Omega', \mathcal{A} \otimes \mathcal{A}')$ by

$$\forall (A,B) \in \mathcal{A} \times \mathcal{A}', \qquad K\mathbb{P}(A,B) \overset{\text{def}}{=} \int_\Omega \mathbb{I}\{\omega \in A\} K(\omega, B) \mathrm{d}\mathbb{P}(\omega)\,.$$

Fixing $t \geq 1$, we get that

$$\mathbb{P}_{\underline{\nu}}^{I_t} = K_{\underline{\nu}}^t \mathbb{P}_{\underline{\nu}}^{I_{t-1}}\,, \qquad \text{where} \quad K_{\underline{\nu}}^t(I_{t-1}, \cdot) = \sum_{a \in [K]} \mathbb{I}\{A_t = a\}\, \nu_a \otimes \mathrm{Leb}_{[0,1]} = \nu_{A_t} \otimes \mathrm{Leb}_{[0,1]} \quad (2.13)$$

is a regular transition kernel, with $\mathrm{Leb}_{[0,1]}$ the Lebesgue measure on $[0,1]$.

Proof. Once again, we rely on optional skipping (see page 10) and consider the sequence of rewards $(X_{a,n})_{a \in [K], n \geq 1}$. Without loss of generality, we assume that $\mathrm{KL}(\nu_a, \zeta_a)$ is finite[7] for each arm $a \in [K]$. As a consequence, note that $\nu_a \ll \zeta_a$ and that $\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a,1})$ is $\mathbb{P}_{\underline{\nu}}$–integrable for all $a \in [K]$.

We will prove that

$$\sum_{a \in [K]} \mathbb{E}_{\underline{\nu}}[N_a(\tau)]\, \mathrm{KL}(\nu_a, \zeta_a) = \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_\tau}, \mathbb{P}_{\underline{\zeta}}^{I_\tau}\Big)\,. \quad (2.14)$$

The result will follow by combining this equality with an application of the data-processing inequality (2.11): $\mathbb{I}\{E\}$ being $\mathcal{F}_\tau$–measurable, it is a random variable depending only on I_τ, hence:

$$\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_\tau}, \mathbb{P}_{\underline{\zeta}}^{I_\tau}\Big) \geq \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{\mathbb{I}\{E\}}, \mathbb{P}_{\underline{\zeta}}^{\mathbb{I}\{E\}}\Big) = \mathrm{KL}\Big(\mathrm{Ber}(\mathbb{P}_{\underline{\nu}}(E)), \mathrm{Ber}(\mathbb{P}_{\underline{\zeta}}(E))\Big)\,.$$

We now prove Equation (2.14). We define $M_0 = 0$ and

$$\forall t \geq 1, \qquad M_t = \log \frac{\mathrm{d}\mathbb{P}_{\underline{\nu}}^{I_t}}{\mathrm{d}\mathbb{P}_{\underline{\zeta}}^{I_t}} - \sum_{a \in [K]} N_a(t)\, \mathrm{KL}(\nu_a, \zeta_a)\,.$$

The process $(M_t)_{t \geq 0}$ is adapted to $(\mathcal{F}_t)_{t \geq 0}$. We will show that $(M_t)_{t \geq 0}$ is a martingale by using the following chain rule property (see, e.g., Wainwright, 2019, Exercice 3.2 or Stoltz, 2022, Lecture 5), which is a consequence of the Radon-Nikodym theorem.

[6] Note that the existence of a measurable space on which $\mathbb{P}_{\underline{\nu}}$ is defined is ensured by Kolmogorov's extension theorem.

[7] If $\mathrm{KL}(\nu_a, \zeta_a) = +\infty$, either the left-hand-size of Equation (2.12) is infinite, or $\mathbb{E}_{\underline{\nu}}\big[N_a(\tau)\big] = 0$ so that arm a is almost surely never pulled and the proof still applies.

Proposition 2.5. *Let K, L be two regular transition kernels from $(\Omega, \mathcal{A})$ to $(\Omega', \mathcal{A}')$ and $\mathbb{P}, \mathbb{Q}$ be two probability distributions on $(\Omega, \mathcal{A})$ with $\mathbb{P} \ll \mathbb{Q}$, and assume that*

- *$K(\omega, \cdot) \ll L(\omega, \cdot)$ for $\mathbb{Q}$–almost all $\omega \in \Omega$,*
- *there exists a version of $(\omega, \omega') \mapsto \frac{\mathrm{d}K(\omega, \cdot)}{\mathrm{d}L(\omega, \cdot)}(\omega')$ which is $\mathcal{A} \otimes \mathcal{A}'$–bi-measurable.*

Then $K\mathbb{P} \ll L\mathbb{Q}$ and the following chain rule property for densities applies:

$$\frac{\mathrm{d}K\mathbb{P}}{\mathrm{d}L\mathbb{Q}}(\omega, \omega') = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega) \cdot \frac{\mathrm{d}K(\omega, \cdot)}{\mathrm{d}L(\omega, \cdot)}(\omega') \qquad \text{for } L\mathbb{Q}\text{–almost all } (\omega, \omega').$$

Fix $t \geq 1$. The transition kernels $K = K_{\underline{\mu}}^t$ and $L = K_{\underline{\zeta}}^t$, defined in (2.13), satisfy the hypothesis of the above proposition, as $\nu_a \ll \zeta_a$ for all $a \in [K]$, and as for all $h_{t-1} = (u_0, y_1, \ldots, u_{t-1})$ and y_t, u_t:

$$\frac{\mathrm{d}K_{\underline{\mu}}^t(h_{t-1}, \cdot)}{\mathrm{d}K_{\underline{\zeta}}^t(h_{t-1}, \cdot)}(y_t, u_t) = \sum_{a \in [K]} \mathbb{I}\{A_t(h_{t-1}) = a\} \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(y_t)\,,$$

where $A_t(h_{t-1})$ denotes the arm chosen by the strategy knowing history h_{t-1}. We can apply the proposition with $\mathbb{P} = \mathbb{P}_{\underline{\nu}}^{I_{t-1}}$ and $\mathbb{Q} = \mathbb{P}_{\underline{\zeta}}^{I_{t-1}}$, as one can show that $\mathbb{P} \ll \mathbb{Q}$ by induction. Using (2.13), and recalling that U_t is an auxiliary random variable independent from all other randomization, this entails

$$\begin{aligned} M_t &= \log \frac{\mathrm{d}\mathbb{P}_{\underline{\nu}}^{I_{t-1}}}{\mathrm{d}\mathbb{P}_{\underline{\zeta}}^{I_{t-1}}} + \log \frac{\mathrm{d}K_{\underline{\mu}}^t(I_{t-1}, \cdot)}{\mathrm{d}K_{\underline{\zeta}}^t(I_{t-1}, \cdot)}(Y_t, U_t) - \sum_{a \in [K]} \left(N_a(t-1) + \mathbb{I}\{A_t = a\}\right) \mathrm{KL}(\nu_a, \zeta_a) \\ &= M_{t-1} + \sum_{a \in [K]} \mathbb{I}\{A_t = a\} \left(\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(Y_t) - \mathrm{KL}(\nu_a, \zeta_a)\right) \\ &= M_{t-1} + \sum_{a \in [K]} \mathbb{I}\{A_t = a\} \left(\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a, N_a(t)}) - \mathrm{KL}(\nu_a, \zeta_a)\right). \end{aligned}$$

We deduce from this expression that the sequence $(M_t)_{t \geq 0}$ is a martingale, as for all $t \geq 1$

$$\begin{aligned} \mathbb{E}_{\underline{\nu}}[M_t | \mathcal{F}_{t-1}] &= M_{t-1} + \sum_{a \in [K]} \mathbb{I}\{A_t = a\}\, \mathbb{E}_{\underline{\nu}}\left[\left(\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a, N_a(t)}) - \mathrm{KL}(\nu_a, \zeta_a)\right) | \mathcal{F}_{t-1}\right] \\ &= M_{t-1} + \sum_{a \in [K]} \mathbb{I}\{A_t = a\} \underbrace{\left(\mathbb{E}_{\underline{\nu}}\left[\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a,1})\right] - \mathrm{KL}(\nu_a, \zeta_a)\right)}_{=0} = M_{t-1}\,, \end{aligned}$$

where we used that A_t is $\mathcal{F}_{t-1}$–measurable and that $X_{a, N_a(t)} | \mathcal{F}_{t-1} \sim \nu_a$ has same law as $X_{a,1}$.

Note that τ was assumed to be $\mathbb{P}_{\underline{\nu}}$–integrable and that the conditional expectation of the absolute increments of $(M_t)_{t \geq 0}$ are almost surely bounded: for any $t \geq 1$ we have

$$\begin{aligned} \mathbb{E}_{\underline{\nu}}\Big[|M_t - M_{t-1}| | \mathcal{F}_{t-1}\Big] &\leq \sum_{a \in [K]} \mathbb{I}\{A_t = a\}\, \mathbb{E}_{\underline{\nu}}\left[\left|\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a, N_a(t)}) - \mathrm{KL}(\nu_a, \zeta_a)\right| | \mathcal{F}_{t-1}\right] \\ &\leq \sup_{a \in [K]} \left\{\mathrm{KL}(\nu_a, \zeta_a) + \mathbb{E}_{\underline{\nu}}\left[\left|\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a,1})\right|\right]\right\} < +\infty\,, \end{aligned}$$

where we used that $\log \frac{\mathrm{d}\nu_a}{\mathrm{d}\zeta_a}(X_{a,1})$ is $\mathbb{P}_{\underline{\nu}}$–integrable by assumption.
By Doob's optional stopping theorem (Grimmett and Stirzaker, 2001, Section 12.5), we obtain that

$$\mathbb{E}_{\underline{\nu}}[M_\tau] = \mathbb{E}_{\underline{\nu}}[M_0] = 0\,.$$

Rearranging the terms directly leads to Equation (2.14), which concludes the proof. □

2.2.3. Lower Bound for Exponential Models

Garivier and Kaufmann (2016) used the fundamental inequality (2.12) to derive a lower bound valid for common models of distributions like the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian variables with a fixed variance $\sigma^2 > 0$, or the model $\mathcal{D}_{\mathcal{B}}$ of Bernoulli distributions with means in $(0,1)$, and more generally for any canonical one-parameter exponential family. Before stating the lower bound, we give a quick description of these families.

Canonical one-parameter exponential families. We follow largely the exposition by Lehmann and Casella (1998, Section 1.5); more details, including the proofs of the stated properties, may be found in the monograph by Lehmann and Casella (1998). A (regular) canonical one-parameter exponential family $\mathcal{D}_{\exp}$ is a set of distributions ν_θ indexed by $\theta \in \Theta$, all absolutely continuous with respect to some measure ρ on $\mathbb{R}$, with densities given by

$$x \longmapsto \frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) \stackrel{\text{def}}{=} \exp(\theta x - b(\theta))\,, \tag{2.15}$$

for some smooth enough normalizing function b. More precisely, b is assumed to be twice differentiable. We also assume that Θ is the natural parameter space, i.e., that Θ contains all possible parameters for ρ,

$$\Theta = \left\{\theta \in \mathbb{R} \,:\, \int_{\mathbb{R}} \exp(\theta y)\,\mathrm{d}\rho(y) < +\infty\right\},$$

and that Θ is an open interval (this latter fact is what regularity stands for). A closed-form expression of b is: for all $\theta \in \Theta$,

$$b(\theta) = \log \int_{\mathbb{R}} \mathrm{e}^{\theta y}\,\mathrm{d}\rho(y)\,. \tag{2.16}$$

The derivative b' of b is a continuous function, by assumption, and it may be shown that it is increasing, so that b' is a one-to-one mapping with a continuous inverse $(b')^{-1}$. In addition, it can be seen, by a differentiation under the integral sign, that $\mathrm{E}(\nu_\theta) = b'(\theta)$ for all $\theta \in \Theta$. Therefore, the distributions in $\mathcal{D}_{\exp}$ may be rather parameterized by their expectations. We denote by $\mathcal{M} = b'(\Theta)$ the open interval of the expectations of distributions in $\mathcal{D}_{\exp}$, and let μ_- and μ_+ be its lower and upper ends:

$$\mathcal{M} = (\mu_-, \mu_+)\,.$$

For each $x \in \mathcal{M}$, there exists a unique distribution in $\mathcal{D}_{\exp}$ with expectation x, namely, $\nu_{(b')^{-1}(x)}$.

Example. Bernoulli distributions, binomial distributions, Poisson distributions and Gaussian distributions with common variance $\sigma^2 > 0$ all are canonical one-parameter exponential families.

In this thesis, any canonical one-parameter exponential family will be simply referred to as an *exponential family*, as we will not consider other exponential families.

Kullback-Leibler divergences for $\mathcal{D}_{\exp}$. For an exponential model $\mathcal{D}_{\exp}$, we may also parameterize the Kullback-Leibler divergence function by the expectations: we define, for all $\theta, \theta' \in \Theta$,

$$d(\mathrm{E}(\nu_\theta), \mathrm{E}(\nu_{\theta'})) \stackrel{\text{def}}{=} \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = (\theta - \theta')b'(\theta) - b(\theta) + b(\theta')\,. \tag{2.17}$$

This defines a divergence d which is strictly convex and differentiable on the open set $\mathcal{M} \times \mathcal{M}$. In particular, d is continuous, is such that $d(\mu, \mu') = 0$ if and only if $\mu = \mu'$, and, for all $\mu \in \mathcal{M}$, both $d(\mu, \cdot\,)$ and $d(\,\cdot\,, \mu)$ are decreasing on $(\mu_-, \mu]$, and increasing on $[\mu, \mu_+)$.

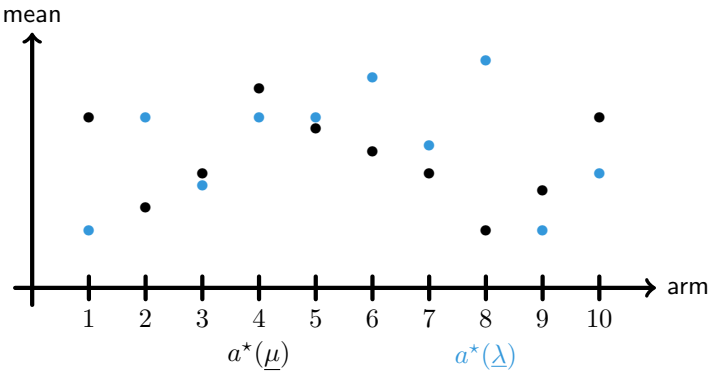


Figure 2.3: A bandit instance $\underline{\mu}$ and an alternative $\underline{\lambda}$.

As we will see, the mean-parametrized Kullback-Leibler divergence of the model of Bernoulli distributions $\mathcal{D}_{\mathcal{B}}$ plays a particular role. To that end, let us denote it by kl. Its closed-form expression reads:

$$\forall p, q \in (0,1), \qquad \mathrm{kl}(p,q) \overset{\text{def}}{=} p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \,.$$

Lower bound for an exponential model. We are now able to state and prove the lower bound obtained by Garivier and Kaufmann (2016) for an exponential model $\mathcal{D}_{\exp}$. As explained in the paragraph on exponential families, distributions are characterized by their means, hence in the sequel we identify $\underline{\nu}$ and its mean vector $\underline{\mu} \in \mathcal{M}^K$; for instance, $\mathbb{P}_{\underline{\mu}} = \mathbb{P}_{\underline{\nu}}$, $\mathbb{E}_{\underline{\mu}} = \mathbb{E}_{\underline{\nu}}$, and $a^\star(\underline{\mu}) = a^\star(\underline{\nu})$.

Before stating the lower bound, we introduce the simplex

$$\Sigma_K \overset{\text{def}}{=} \left\{ \underline{v} \in [0,1]^K \,:\, v_1 + \cdots + v_K = 1 \right\},$$

and the set of *alternative* bandits[8] to a problem $\underline{\mu}$, which are bandit problems with a different optimal arm than $\underline{\mu}$ (see Figure 2.3):

$$\mathrm{Alt}(\underline{\mu}) \overset{\text{def}}{=} \left\{ \underline{\lambda} \text{ in } \mathcal{D}_{\exp} \,:\, a^\star(\underline{\lambda}) \neq a^\star(\underline{\mu}) \right\}. \tag{2.18}$$

Theorem 2.6. **[Garivier and Kaufmann, 2016, Theorem 1]**
Let $\mathcal{D}_{\exp}$ be an exponential model. Let $\delta \in (0, \frac{1}{2})$. For all δ–correct strategy and all bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$ with a unique optimal arm,

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta] \geq T(\underline{\mu})\, \mathrm{kl}(\delta, 1-\delta) \geq T(\underline{\mu}) \log \frac{1}{2.4\delta} \,, \tag{2.19}$$

where $T(\underline{\mu})$ is the characteristic time *of $\underline{\mu}$, defined as*

$$T(\underline{\mu})^{-1} \overset{\text{def}}{=} \sup_{\underline{v} \in \Sigma_K} \inf_{\underline{\lambda} \in \mathrm{Alt}(\underline{\mu})} \sum_{a \in [K]} v_a \, d(\mu_a, \lambda_a) \,. \tag{2.20}$$

Example. For the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian variables with variance $\sigma^2 > 0$, the Kullback-Leibler divergence enjoys the simple closed-form expression

$$\forall \mu, \mu' \in \mathbb{R}, \qquad d(\mu, \mu') = \frac{(\mu' - \mu)^2}{2\sigma^2} \,, \tag{2.21}$$

[8] Again, alternative bandits $\underline{\zeta}$ will be parameterized by their expectation $\underline{\lambda}$.

from which it can be proved that $T(\underline{\mu})$ satisfies the following inequalities:

$$\sum_{a\neq a^\star}\frac{1}{\Delta_a^2}\leq\sum_{a\neq a^\star}\frac{1}{\Delta_a^2}+\frac{1}{\Delta_2^2}\leq\frac{T(\underline{\mu})}{2\sigma^2}\leq 2\Big(\sum_{a\neq a^\star}\frac{1}{\Delta_a^2}+\frac{1}{\Delta_2^2}\Big)\leq 4\sum_{a\neq a^\star}\frac{1}{\Delta_a^2}\,.$$

This should be compared to the high probability upper bound (2.10) obtained by Even-Dar et al. (2006) with an adaptive elimination algorithm, which applies up to a variance term for the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ and gets an additional factor $\log\frac{K}{\Delta_{\min}}$.

Proof. Let $\underline{\mu}$ be a K–armed bandit problem with a unique optimal arm, and fix an alternative bandit instance $\underline{\lambda}\in\mathrm{Alt}(\underline{\mu})$. If $\mathbb{E}_{\underline{\mu}}[\tau_\delta]=+\infty$ the result holds. We assume in the rest of the proof that $\mathbb{E}_{\underline{\mu}}[\tau_\delta]<+\infty$, and, as a consequence, $\tau_\delta<+\infty$ $\mathbb{P}_{\underline{\mu}}$–a.s..

Considering $E=\{\tau_\delta=+\infty \text{ or } \hat{a}_{\tau_\delta}\neq a^\star(\underline{\mu})\}\in\mathcal{F}_{\tau_\delta}$, and recalling that the strategy is δ–correct, we get, on the one hand, as $\tau_\delta<+\infty$ $\mathbb{P}_{\underline{\mu}}$–a.s.,

$$\mathbb{P}_{\underline{\mu}}(E)=\mathbb{P}_{\underline{\mu}}(\tau_\delta=+\infty \text{ or } \hat{a}_{\tau_\delta})=\mathbb{P}_{\underline{\mu}}(\tau_\delta<+\infty,\hat{a}_{\tau_\delta}\neq a^\star(\underline{\mu}))\leq\delta\,,$$

and on the other hand, as $a^\star(\underline{\mu})\neq a^\star(\underline{\lambda})$,

$$\mathbb{P}_{\underline{\lambda}}(E)=\mathbb{P}_{\underline{\lambda}}(\tau_\delta=+\infty \text{ or } \hat{a}_{\tau_\delta}\neq a^\star(\underline{\mu}))\geq\mathbb{P}_{\underline{\lambda}}(\tau_\delta=+\infty \text{ or } \hat{a}_{\tau_\delta}=a^\star(\underline{\lambda}))\geq 1-\delta\,.$$

Injecting those two inequalities in the fundamental inequality (2.12), and using the monotonicity properties, recalled after Equation (2.17), of the kl divergence with $\delta<\frac{1}{2}$, we obtain

$$\begin{aligned}\mathbb{E}_{\underline{\mu}}[\tau_\delta]\times\sum_{a\in[K]}\frac{\mathbb{E}_{\underline{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}\,d(\mu_a,\lambda_a)&=\sum_{a\in[K]}\mathbb{E}_{\underline{\mu}}[N_a(\tau_\delta)]\,d(\mu_a,\lambda_a)\\&\geq\mathrm{kl}\Big(\mathbb{P}_{\underline{\mu}}(\hat{a}_{\tau_\delta}\neq a^\star(\underline{\mu})),\,\mathbb{P}_{\underline{\lambda}}(\hat{a}_{\tau_\delta}\neq a^\star(\underline{\mu}))\Big)\\&\geq\mathrm{kl}(\delta,1-\delta)\,.\end{aligned}$$

This inequality holds for any alternative bandit $\underline{\lambda}$, hence, taking the infimum over $\mathrm{Alt}(\underline{\mu})$ entails

$$\begin{aligned}\mathrm{kl}(\delta,1-\delta)&\leq\mathbb{E}_{\underline{\mu}}[\tau_\delta]\times\inf_{\underline{\lambda}\in\mathrm{Alt}(\underline{\mu})}\sum_{a\in[K]}\frac{\mathbb{E}_{\underline{\mu}}[N_a(\tau_\delta)]}{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}\,d(\mu_a,\lambda_a)\\&\leq\mathbb{E}_{\underline{\mu}}[\tau_\delta]\times\sup_{\underline{v}\in\Sigma_K}\inf_{\underline{\lambda}\in\mathrm{Alt}(\underline{\mu})}\sum_{a\in[K]}v_a\,d(\mu_a,\lambda_a)\,,\end{aligned}\tag{2.22}$$

using in the last inequality that $\tau_\delta=\sum_{a\in[K]}N_a(\tau_\delta)$. This concludes the proof of the first inequality, while the second can be derived by a mere calculus. □

2.2.4. The Sample Complexity Optimization Problem

As in the previous subsection, we consider here an exponential model $\mathcal{D}_{\exp}$. As we will see, the solution of the optimization problem (2.20) defining $T(\underline{\mu})$ is of high interest to design good strategies.

Transportation costs. Given a fixed proportion $\underline{v}\in\Sigma_K$, reaching the infimum over $\mathrm{Alt}(\underline{\mu})$ in optimization problem (2.20) is done by considering bandit instances for which only two arms differ from $\underline{\mu}$, namely the best arm and one of the sub-optimal arms (see Figure 2.4):

$$\begin{aligned}\inf_{\underline{\lambda}\in\mathrm{Alt}(\underline{\mu})}\sum_{a\in[K]}v_a\,d(\mu_a,\lambda_a)&=\min_{a\neq a^\star}\inf_{\substack{\lambda_{a^\star},\lambda_a\in\mathcal{M}\\ \lambda_{a^\star}<\lambda_a}}v_{a^\star}\,d(\mu_{a^\star},\lambda_{a^\star})+v_a\,d(\mu_a,\lambda_a)\\&=\min_{a\neq a^\star}\inf_{x\in[\mu_a,\mu^\star]}v_{a^\star}\,d(\mu^\star,x)+v_a\,d(\mu_a,x)\,,\end{aligned}\tag{2.23}$$

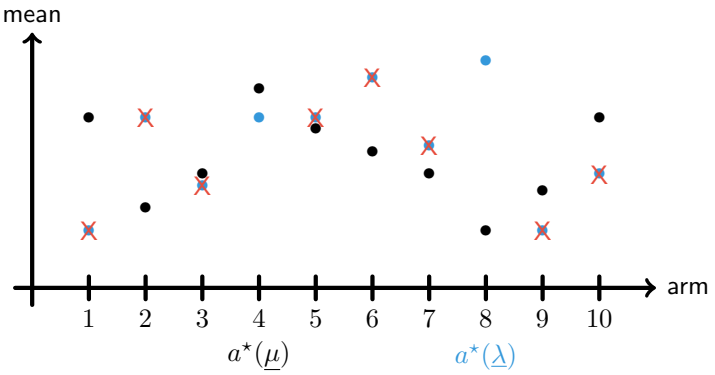


Figure 2.4: To decrease its transportation cost with $\underline{\mu}$, an alternative bandit problem $\underline{\lambda}$ should only move two arms, namely the optimal arm of $\underline{\mu}$ and one sub-optimal arm of $\underline{\mu}$.

by using the continuity and monotonicity properties of d, where $a^\star = a^\star(\underline{\mu})$ and we recall that $\mu^\star = \mu_{a^\star}$ is the optimal mean of $\underline{\mu}$. The quantity

$$\mathrm{TC}_{a\to a^\star}(\underline{\mu}, \underline{v}) \stackrel{\mathrm{def}}{=} \inf_{x\in[\mu_a,\mu^\star]} v_{a^\star}\, d(\mu^\star, x) + v_a\, d(\mu_a, x)$$

can be seen as a *transportation cost* representing the difficulty of changing the distributions of arms a and $a^\star$ so as to make arm a optimal, given the pulling frequencies $\underline{v}$. The monotonicity properties of d ensures that the continuous function $x \mapsto v_{a^\star}\, d(\mu^\star, x) + v_a\, d(\mu_a, x)$ reaches is infimum in $[\mu_a, \mu^\star]$, and using the closed-form (2.17) of d, we see that its derivative vanishes if and only if

$$v_{a^\star}(x - \mu^\star) + v_a(x - \mu_a) = 0\,,$$

or, equivalently, if and only if $x = \overline{\mu}_{a^\star, a, \underline{v}}$, where

$$\overline{\mu}_{a^\star, a, \underline{v}} \stackrel{\mathrm{def}}{=} \frac{v_{a^\star}\mu^\star + v_a\mu_a}{v_{a^\star} + v_a}\,,$$

so that

$$\mathrm{TC}_{a\to a^\star}(\underline{\mu}, \underline{v}) = v_{a^\star} d(\mu^\star, \overline{\mu}_{a^\star, a, \underline{v}}) + v_a d(\mu_a, \overline{\mu}_{a^\star, a, \underline{v}})\,.$$

Hence we proved that optimization problem (2.20) rewrites

$$T(\underline{\mu})^{-1} = \sup_{\underline{v}\in\Sigma_K} \min_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu}, \underline{v}) = \sup_{\underline{v}\in\Sigma_K} \min_{a\neq a^\star} v_{a^\star}\, d(\mu^\star, \overline{\mu}_{a^\star, a, \underline{v}}) + v_a\, d(\mu_a, \overline{\mu}_{a^\star, a, \underline{v}})\,. \tag{2.24}$$

Remark 2.7. When considering a general model $\mathcal{D}$ (even non-parametric), the transportation cost becomes

$$\mathrm{TC}_{a\to a^\star}(\underline{\nu}, \underline{v}) \stackrel{\mathrm{def}}{=} \inf_{x\in[\mu_a,\mu^\star]} v_{a^\star}\, \mathcal{K}^<_{\inf}(\nu^\star, x) + v_a\, \mathcal{K}^>_{\inf}(\nu_a, x)\,,$$

where we define, for $\nu \in \mathcal{D}$ and $x \in \mathbb{R}$,

$$\mathcal{K}^<_{\inf}(\nu, x) \stackrel{\mathrm{def}}{=} \inf\{\mathrm{KL}(\zeta, \nu) : \zeta \in \mathcal{D} \ \text{ s.t. } \ \mathrm{E}(\zeta) < x\}\,,$$
$$\text{and} \qquad \mathcal{K}^>_{\inf}(\nu, x) \stackrel{\mathrm{def}}{=} \inf\{\mathrm{KL}(\zeta, \nu) : \zeta \in \mathcal{D} \ \text{ s.t. } \ \mathrm{E}(\zeta) > x\}\,.$$

A generalization of Theorem 2.6 to any model $\mathcal{D}$ involving those costs is straightforward (see Agrawal et al., 2020).

An optimal weight vector. Garivier and Kaufmann (2016) proved that the sample complexity optimization problem (2.20) admits a unique maximizer $\underline{w}(\underline{\mu})$, called the *optimal weight vector* of $\underline{\mu}$:

$$\left\{\underline{w}(\underline{\mu})\right\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a\, d(\mu_a,\lambda_a) = \operatorname*{argmax}_{\underline{v}\in\Sigma_K} \min_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})\,.$$

The main arguments are summarized below[9].

- Existence: it can be seen that $\underline{v}\mapsto \min_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})$ is a continuous function since each $\underline{v}\mapsto \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})$ is continuous. Hence, it admits a maximum on the compact set Σ_K.
- Unicity:
  1. First, we prove that, given a fixed proportion $w_{a^\star}\in(0,1)$, there exists a unique vector $\underline{w}\in\Sigma_K$ which equalizes all transportation costs, i.e., such that
  $$\exists y\in\mathbb{R}_+, \forall a\neq a^\star, \qquad \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{w}) = y\,.$$
  This result is a consequence of both the constraint $\underline{w}\in\Sigma_K$ and the fact that, given a fixed $v_{a^\star}$, each cost $\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})$ is an increasing and continuous function of v_a.
  2. Then, we remark that a maximizer $\underline{w}$ has to equalize the transportation costs. Let us assume that this does not hold, i.e., that the set $\mathcal{A}_{\min} = \operatorname{argmin}_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{w})$ does not contain all sub-optimal arms. Let us transform $\underline{w}$ into a new vector $\underline{w}'\in\Sigma_K$ such that $w'_a > w_a$ for each $a\in\mathcal{A}_{\min}$, $w_{a^\star}=w'_{a^\star}$, and $w'_a<w_a$ otherwise. If the weights are modified slightly enough, this entails[10] that the minimal costs increase while the other costs only decrease sufficiently low, i.e., that
  $$\min_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{w}) < \min_{a\neq a^\star} \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{w}')\,,$$
  or, in other words, that $\underline{w}$ is not a maximizer, which is a contradiction.
  3. From the two first points, potential maximizers belong to a set $\{\underline{w}^\beta : \beta\in(0,1)\}$ of weight vectors that are parameterized by the proportion $\beta = w^\beta_{a^\star}$ associated to the optimal weight. By defining $y(\beta)$ as the (common) value of the transportation costs under proportions $\underline{w}^\beta$, it might be seen that $y(\beta)$ is a differentiable function, such that $y(\beta)\to 0$ when $\beta\to 0$ and $\beta\to 1$, and whose derivative vanishes only once, at some $\beta^\star\in(0,1)$. Hence the optimal maximizer is unique and equal to $\underline{w}^{\beta^\star}$.

Note that the proof of unicity indicates that the optimal weight vector $\underline{w}(\underline{\mu})$ equalizes the transportation costs:

$$\forall a\neq a^\star, \qquad \mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{w}(\underline{\mu})) = T(\underline{\mu})^{-1}\,.$$

If we want a strategy for which inequality (2.22) is an equality, then its average proportions of draws $\left(\mathbb{E}_{\underline{\mu}}[N_a(\tau_\delta)]/\mathbb{E}_{\underline{\mu}}[\tau_\delta]\right)_{a\in[K]}$ should exactly be $\underline{w}(\underline{\mu})$. To put it differently, if those average proportions are not close to $\underline{w}(\underline{\mu})$, then the performance of the strategy will be far from lower bound (2.19).

Computability of the optimal weight vector. Garivier and Kaufmann (2016) proved that solving the optimization problem (2.20) reduces to determining the root of a one-variable increasing function. By applying a bisection method, it is then possible to compute $\underline{w}(\underline{\mu})$ with arbitrary precision. We will denote by `Optimal-Weights` an algorithm computing this optimal vector.

Using their procedure, they also proved some regularity results concerning the solution of $\underline{w}(\underline{\mu})$:

- the function $\underline{\mu}\mapsto\underline{w}(\underline{\mu})$ is continuous (at problems having a unique optimal arm),
- all arms have to be linearly pulled: $w_{\min}(\underline{\mu}) \stackrel{\text{def}}{=} \min_{a\in[K]} w_a(\underline{\mu}) > 0$,

[9] Note, however, that the structure of the proof does not exactly follow the path given here, due to technicalities (see Garivier and Kaufmann, 2016, Appendix A.2).

[10] It is, again, a consequence of the increasing and continuity properties of each $\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})$ relatively to v_a.

- if arms are ordered so that $\mu_{(1)} > \mu_{(2)} \geq \ldots \geq \mu_{(K)}$, then $w_{(2)}(\underline{\mu}) \geq \ldots \geq w_{(K)}(\underline{\mu})$, but there may be instances (e.g., for Bernoulli instances) for which $w_{(1)}(\underline{\mu}) < w_{(2)}(\underline{\mu})$.

Remark 2.8. It is easy to find instances for which $w_{(1)}(\underline{\mu}) \gg w_{(2)}(\underline{\mu})$. As a consequence, for such instances, elimination algorithms presented in Section 2.2.1 cannot get close to the lower bound (2.19), as the last two remaining arms are pulled equally often.

Contribution. We make progress in the understanding of the solution of optimization problem (2.20) in two directions: first, in Chapter 3, we obtain *regularity results* for the specific case of a Gaussian model, and then, in Section 6.2 we characterize, for all exponential models, the solution $\underline{w}(\underline{\mu})$ as the unique fixed point of some transformation. Those theoretical contributions might be used to improve fixed-confidence best-arm identification algorithms, as we will explain in Contributions 2.13 and 2.14.

The solution for a Gaussian model. Considering the specific model $\mathcal{D}_{\mathcal{N}_1}$ of standard Gaussian distributions (or more generally any model of Gaussian distributions with common variance $\sigma^2 > 0$ and bounded means), we will present in Chapter 3 a new method for computing the optimal weight vector $\underline{w}(\underline{\mu})$. The proposed procedure speeds up the resolution of the computation by using Newton's method on a convex function. In addition, we obtain new quantitative regularity results concerning the solution of optimization problem (2.20):

- First, we prove that $a^\star = \operatorname{argmax}_{a\in[K]} w_a(\underline{\mu})$ for this model, or equivalently $w_{(1)}(\underline{\mu}) > w_{(2)}(\underline{\mu})$. Optimal weights are thus ordered as the means for a Gaussian model,
- Then, we derive the following bounds for $w_{a^\star}(\underline{\mu})$ and $T(\underline{\mu})$:

$$\frac{1}{1+\sqrt{K-1}} \leq w_{a^\star}(\underline{\mu}) \leq \frac{1}{2}\,,$$
$$\max\left(\frac{8}{\Delta_{\min}^2}, 4\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\right) \leq T(\underline{\mu}) \leq 2\frac{\left(1+\sqrt{K-1}\right)^2}{\Delta_{\min}^2}\,,$$

  where $\overline{\Delta^2} \overset{\text{def}}{=} \frac{1}{K-1}\sum_{a\neq a^\star}\Delta_a^2$ is the average squared gap and $\Delta_{\min}$ has been defined in page 39. For each inequality, we characterize instances for which the equality holds, which shows the tightness of those bounds in all generality.
- Also, we study the variations of $\underline{w}(\underline{\mu})$ and $T(\underline{\mu})$ when moving one or several arms of $\underline{\mu}$. For instance (see Figure 2.5), increasing the mean of a sub-optimal arm will increase its associated optimal weight and increase the characteristic time $T(\underline{\mu})$.
- Finally, we prove that both $\underline{w}$ and T are locally Lipschitz functions, giving a quantification of the continuity result obtained by Garivier and Kaufmann (2016).

Remark. All those results are specific to the Gaussian case. In particular, similar regularity results of $\underline{w}$ and T for other models are still to be determined.

A fixed point property for exponential models. For a general exponential model $\mathcal{D}_{\exp}$, we use sufficient and necessary conditions satisfied by the optimal weight vector $\underline{w}(\underline{\mu})$ to prove in Section 6.2 that $\underline{w}(\underline{\mu})$ is the unique fixed point of a transformation $\underline{W} : \operatorname{int}(\Sigma_K) \to \operatorname{int}(\Sigma_K)$, which roughly speaking modifies a weight vector so as to decrease the difference between the transportation costs. This could entail the design of new empirical procedures that speed up the computation of $\underline{w}(\underline{\mu})$.

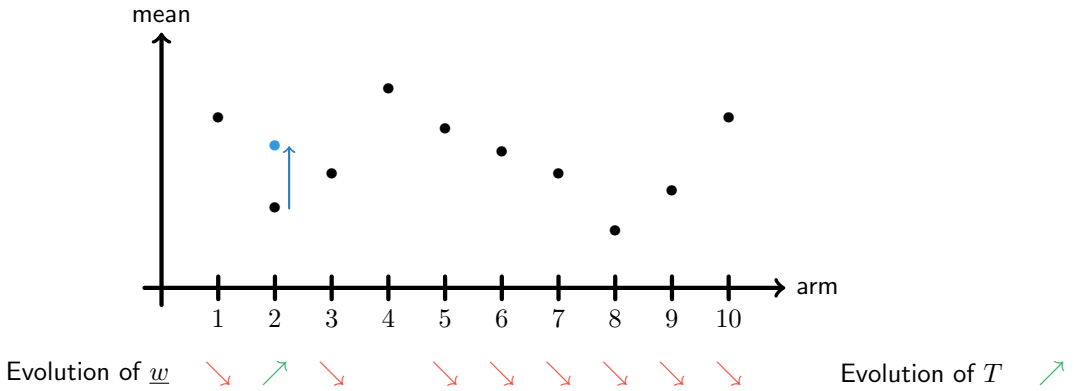


Figure 2.5: By increasing the mean of a sub-optimal arm, its associated optimal weight increases while those of other sub-optimal arms decrease, and the characteristic time increases.

2.2.5. The `Track-and-Stop` Algorithm: an Asymptotically Optimal Strategy

To obtain strategies that could reach lower bound (2.19), we have seen that we have no other choice than finding sampling rules which ensure that the empirical proportions of draws are close to $\underline{w}(\underline{\mu})$. Of course, the learner does not know $\underline{\mu}$, so she cannot directly compute $\underline{w}(\underline{\mu})$. However, by continuity of $\underline{w}$, a strategy can estimate it by using the current empirical estimate $\hat{\underline{\mu}}(t)$ available at step t. This is the main idea of the `Track-and-Stop` algorithm that we will now describe.

Sampling rule. At the beginning of time step t, we have access to $\hat{\underline{\mu}}(t-1)$ the maximum likelihood estimator of $\underline{\mu}$. Hence, using the `Optimal-Weights` procedure, we can compute the plug-in estimate $\hat{\underline{w}}(t-1) \overset{\text{def}}{=} \underline{w}(\hat{\underline{\mu}}(t-1))$ which is a good estimate of $\underline{w}(\underline{\mu})$ when $\hat{\underline{\mu}}(t)$ is close to $\underline{\mu}$, thanks to the continuity of $\underline{w}$. The sampling rule consists in *tracking* this proportion $\hat{\underline{w}}(t-1)$: it compares $\hat{\underline{w}}(t-1)$ with the empirical proportions $\frac{N(t-1)}{(t-1)}$ and chooses the arm which has been most under-sampled. There are two approaches:

- we can work in a cumulative way, by tracking the average $\frac{1}{t-1}\sum_{s\in[t-1]} \hat{\underline{w}}(s)$. This option, referred to as *C-tracking*, is the simplest to analyze theoretically as the variability between the two consecutive tracked weights is low,
- or we can use a direct approach by tracking the current weight $\hat{\underline{w}}(t-1)$. This option called *D-tracking* performs empirically better than C-tracking.

Unfortunately, due to the randomness, it might happen that some good arms with poor initial observations stay under-sampled and hence do not improve their estimates, which makes the strategy inefficient. To counter this issue[11], the strategy forces exploration by observing the least pulled arm among the set $U(t-1)$ of arms that are under-sampled with respect to the sub-linear rate $\sqrt{t}$:

$$U(t-1) \overset{\text{def}}{=} \left\{ a \in [K] \,:\, N_a(t-1) < \sqrt{t-1} - \frac{K}{2} \right\}.$$

The sub-linear rate $\sqrt{t}$ is in fact arbitrary and ensures guarantees on the minimal convergence rate of the empirical mean to $\underline{\mu}$ (see Section 6.3.3). The sampling rule of `Track-and-Stop` is presented in Algorithm 4.

Remark. Note that we will present, in Contribution 2.13, a new sampling rule which modifies the tracking of `Track-and-Stop` in a natural way and which enjoys the same exploration rate $\sqrt{t}$ without any force exploration mechanism.

[11] We present here the way forced exploration is ensured with D-tracking. For C-tracking, the mechanism is different but we omit its presentation for the sake of simplicity.

Algorithm 4: `Track-and-Stop` sampling rule at step $t > K$

Input: history of observations I_{t-1}
Output: next arm to observe A_t

$U(t-1) \leftarrow \{a \in [K] : N_a(t-1) < \sqrt{t-1} - \frac{K}{2}\}$
$\underline{\hat{w}}(t-1) \leftarrow$ `Optimal-Weights`$(\underline{\hat{\mu}}(t-1))$
if $U(t-1) \neq \varnothing$ **then**
  $A_t \leftarrow \operatorname{argmin}_{a \in U(t-1)} N_a(t-1)$
else
  /* C-tracking */
  Choose $A_t \in \operatorname{argmin}_{a \in [K]} N_a(t-1) - \sum_{s \in [t-1]} \hat{w}_a(s)$
  /* D-tracking */
  Choose $A_t \in \operatorname{argmin}_{a \in [K]} N_a(t-1) - (t-1)\hat{w}_a(t-1)$

Stopping condition. Given observations up to time t, the algorithm should stop if it has statistical evidence that it found the optimal arm with risk δ. The problem can be seen as a statistical test. Garivier and Kaufmann (2016) introduced the generalized log–likelihood ratio (see Chernoff, 1959) between arms a and b:

$$Z_{a,b}(t) \stackrel{\text{def}}{=} \log \frac{\sup_{\substack{\lambda_a, \lambda_b \in \mathcal{M} \\ \lambda_a \geq \lambda_b}} \prod_{s=1}^{N_a(t)} p_{\lambda_a}(X_{a,s}) \prod_{s=1}^{N_b(t)} p_{\lambda_b}(X_{b,s})}{\sup_{\substack{\lambda_a, \lambda_b \in \mathcal{M} \\ \lambda_a \leq \lambda_b}} \prod_{s=1}^{N_a(t)} p_{\lambda_a}(X_{a,s}) \prod_{s=1}^{N_b(t)} p_{\lambda_b}(X_{b,s})},$$

where $p_\lambda = \frac{\mathrm{d}\zeta}{\mathrm{d}\rho}$ if ζ is the distribution of $\mathcal{D}_{\exp}$ of mean λ, and whose expression is given in (2.15). It may be shown that $Z_{a,b}(t)$ is the empirical transportation cost from arm b to arm a up to a factor t and the sign of $\hat{\mu}_a(t) - \hat{\mu}_b(t)$:

$$Z_{a,b}(t) = t\, \mathrm{TC}_{b \to a}\Big(\underline{\hat{\mu}}(t), \tfrac{\underline{N}(t)}{t}\Big) \operatorname{sgn}\Big(\hat{\mu}_a(t) - \hat{\mu}_b(t)\Big)\,.$$

If $Z_{a,b}(t)$ is large enough then we can confidently reject H_0: $\mu_a \leq \mu_b$ and conclude to H_1: $\mu_a > \mu_b$. When considering all arms, we define[12]

$$\begin{aligned} Z(t) &\stackrel{\text{def}}{=} \max_{a \in [K]} \min_{b \neq a} Z_{a,b}(t) \\ &= \min_{b \neq a^\star(\underline{\hat{\mu}}(t))} Z_{a^\star(\underline{\hat{\mu}}(t)),b}(t) \\ &= t \min_{b \neq a^\star(\underline{\hat{\mu}}(t))} \mathrm{TC}_{b \to a^\star(\underline{\hat{\mu}}(t))}\Big(\underline{\hat{\mu}}(t), \frac{\underline{N}(t)}{t}\Big) \qquad (2.25) \\ &= \inf_{\underline{\lambda} \in \mathrm{Alt}(\underline{\hat{\mu}}(t))} \sum_{a \in [K]} N_a(t)\, d(\mu_a, \lambda_a)\,. \qquad (2.26) \end{aligned}$$

The `Global-Likelihood-Ratio` stopping rule (or *Chernoff stopping rule*), see Algorithm 5, associated to a given threshold function $\beta(t, \delta)$ consists in stopping the strategy as soon as $Z(t)$ exceeds

[12] Where $a^\star(\underline{\hat{\mu}}(t))$ is any of the optimal arms of $\underline{\hat{\mu}}(t)$.

Algorithm 5: `Global-Likelihood-Ratio` stopping rule at step $t > K$

Input: history of observations I_t
threshold function $\beta(t, \delta)$

$Z(t) \leftarrow \max_{a \in [K]} \min_{b \neq a} Z_{a,b}(t)$ // $Z_{a,b}(t)$ is defined in (4.7)
if $Z(t) > \beta(t, \delta)$ **then**
  Stop
else
  Continue

$\beta(t, \delta)$:

$$\tau_\delta \stackrel{\text{def}}{=} \inf\Big\{ t \geq 1 \,:\, Z(t) > \beta(t, \delta) \Big\} .$$

Of course, the associated decision rule will be to return the best empirical arm at time τ_δ (see the `Empirical-Best` Procedure 2). The stopping rule problem now reduces to the research of threshold functions which ensure δ–correctness, a statistical problem that does not depend on the sampling rule. Using Equation (2.26), we get $Z(t) \leq \frac{t}{T(\hat{\mu}(t))}$ and expect $Z(t) \sim \frac{t}{T(\underline{\mu})}$ for large values of t. For a time-independent threshold $\beta(t, \delta) = \beta_\delta$, this indicates that the strategy should stop after approximately $\tau_\delta \sim T(\underline{\mu})\beta_\delta$ steps, and the δ–correctness condition might force $\beta_\delta \gtrsim \log \frac{1}{\delta}$ at the sight of lower bound (2.19).

Obtaining δ–correctness is based on time-uniform bounds that require time-dependent thresholds. Garivier and Kaufmann (2016) proposed the following threshold using deviation results.

Theorem 2.9. [Garivier and Kaufmann, 2016, Proposition 12]
Consider an exponential model $\mathcal{D}_{\text{exp}}$. Let $\delta \in (0, 1)$ and $\alpha > 1$. There exists a constant $R = R(\alpha, K)$ such that, whatever the sampling rule, using the `Global-Likelihood-Ratio` *stopping rule (Algorithm 5) with threshold*

$$\beta(t, \delta) \stackrel{\text{def}}{=} \log \frac{R t^\alpha}{\delta} , \tag{2.27}$$

and the `Empirical-Best` *decision rule (Algorithm 2) ensures that the strategy is δ–correct.*

The performance of a strategy highly depends on both its sampling and stopping rules. While many efficient sampling rules have been studied in the literature (see Section 2.2.7), the use of the `Global-Likelihood-Ratio` stopping rule appears to be consensual: it does not seem that it might be significantly improved, apart from slightly reducing the threshold. In the remainder of this chapter, except for elimination strategies, all considered fixed-confidence best-arm identification strategies will use the `Global-Likelihood-Ratio` stopping rule with threshold (2.27) for a given value of α and the `Empirical-Best` decision rule. The name of the strategy will be defined as the name of the sampling rule.

Remark.
- With threshold (2.27), the stopping time τ_δ is almost surely finite. Indeed, intuitively, for t large enough and under a good sampling rule, based on Equation (2.25) we get

$$Z(t) \simeq \frac{t}{T(\underline{\mu})} ,$$

  hence $Z(t)$ grows much faster than $\beta(t, \delta)$.
- Obtaining δ–correctness with threshold (2.27) crucially depends on a condition on R that does not scale pretty well with the number of arms K. However, this is not an issue numerically:

more convenient thresholds, which are closer to $\log \frac{1}{\delta}$ like

$$\beta(t,\delta) = \log \frac{1+\log t}{\delta}\,,$$

are known to be empirically δ–correct (although there is no theoretical proof of this observation).
More generally, theoretical algorithms are often adapted in experiments to versions that perform better in practice (see, e.g., the use of D-tracking instead of C-tracking for the stopping rule of `Track-and-Stop`).

Open question. Finding theoretical arguments to obtain better thresholds — as close to $\log \frac{1}{\delta}$ as possible —, is an important statistical question, whose answer will get implications beyond bandit problems. Recent techniques using mixture martingales (see Garivier and Kaufmann, 2021; Kaufmann and Koolen, 2021 and very recently Chowdhury et al., 2023) proposed thresholds with better asymptotical behaviors which are however model-dependent and quite less explicit.

Asymptotic optimality. The `Track-and-Stop` sampling rule has been chosen so that the empirical frequencies of pulls converge to $\underline{w}(\underline{\mu})$, while the design of the `Global-Likelihood-Ratio` stopping-rule with threshold (2.27) ensures the δ–correctness. We may wonder how close the performance of the corresponding strategy is from lower bound (2.19). It turns out that, in the *asymptotic* regime when δ goes to 0, the upper bound of `Track-and-Stop` matches the lower bound.

Theorem 2.10. *Consider an exponential model $\mathcal{D}_{\exp}$. The* `Track-and-Stop` *strategy, with threshold (2.27) for a fixed $\alpha > 1$, satisfies, for all bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$ with a unique optimal arm,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \le T(\underline{\mu})\,.$$

Remark 2.11. The original result of Garivier and Kaufmann (2016, Theorem 14) only ensures that

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \le \alpha\, T(\underline{\mu})\,.$$

The multiplicative factor α can be avoided using recent proof techniques, as we explain in Section 6.4.

This result indicates that the lower bound (2.19) cannot be improved in the asymptotic regime. The `Track-and-Stop` strategy is hence *asymptotically optimal* for all exponential models.

Definition 2.12. [asymptotically optimal strategy]
A strategy is said to be *asymptotically optimal* on a model $\mathcal{D}$ if for all bandit problems $\underline{\mu}$ in $\mathcal{D}$ with a unique optimal arm,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \le T(\underline{\mu})\,.$$

Empirical performance. The `Track-and-Stop` algorithm is not only a theoretical contribution, but it also proved to be numerically efficient: the strategy outperforms its competitors in a wide variety of settings.

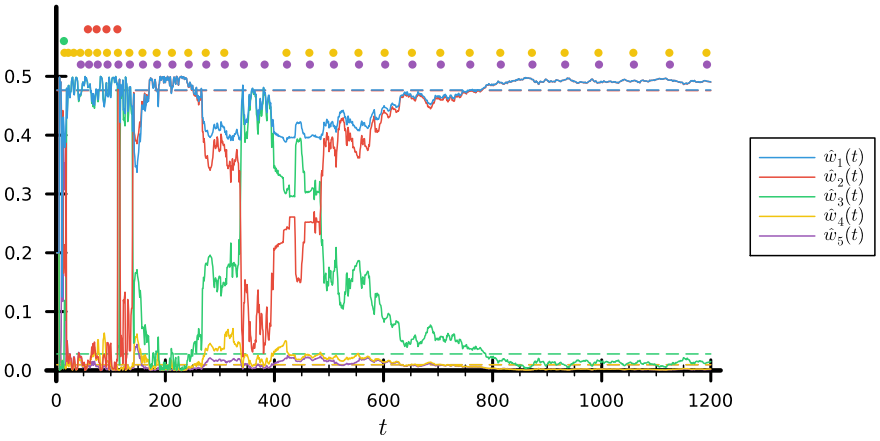


Figure 2.6: Instability of the targeted weights $\hat{\underline{w}}(t)$ during the first 1200 rounds of a simulation of `Track-and-Stop` with parameters $\delta = 0.01$, $\underline{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$ and D-tracking. Dots correspond to the use of forced exploration for under-sampled arms, while the values of the optimal weights $\underline{w}(\underline{\mu}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ are dashed.

2.2.6. Non-Asymptotic Guarantees

Some questions raised by `Track-and-Stop`. The `Track-and-Stop` strategy suffers from certain shortcomings illustrated in Figure 2.6:

- first, the sampling rule appears to be pretty unstable, especially at the beginning: the target frequencies can vary significantly as the estimated means fluctuate before stabilizing around their expectations,
- second, `Track-and-Stop` does not exhibit the intuitively desirable behavior to sample uniformly in the beginning, until sufficient information has been gathered for significant differences between the arms to emerge. This is in contrast with elimination strategies, which are sub-optimal but intuitively appealing.
- also, the forced exploration appears very arbitrary, with a rate of $\sqrt{t}$ that has no other justification than lying somewhere between constant and linear functions.

Altogether, these issues lead for example to unpredictable and irregular behaviors at the beginning of multiple A/B testing cases with many arms extremely close to optimal.

Non-asymptotic guarantees.

Open question. As we have seen, `Track-and-Stop` is asymptotically optimal. The next natural direction is to obtain guarantees in the *non-asymptotic* regime, for fixed values of δ. A close look into the proofs of Garivier and Kaufmann (2016) shows that the theoretical guarantees proved so far are asymptotic in nature. Even if the strategy works pretty well empirically, one needs to develop different proof techniques to get non-asymptotic guarantees.

Note that lower bound (2.19) might not be well-suited for the moderate regime (values of δ that are not too small): Simchowitz et al. (2017) proved a lower bound for moderate values of δ that does not depend on the risk δ. This moderate setting has led to strategies that are sub-optimal by a multiplicative constant but are proved to satisfy explicit non-asymptotic bounds (Karnin et al., 2013; Jamieson et al., 2014; Chen et al., 2017). Yet, can we design asymptotically optimal strategies with finite risk bounds that match the asymptotic complexity?

Contribution 2.13. Due to the instability of the tracking procedure of `Track-and-Stop` that we described in the previous paragraph, it seems difficult to obtain theoretical non-asymptotic bounds. In order to address this problem, we present a novel algorithm named `Exploration-Biased-Sampling` in Chapter 4. This algorithm effectively resolves all the limitations discussed earlier and attains certain non-asymptotic assurances, specifically for a Gaussian model with a shared variance $\sigma^2 > 0$ and constrained means.

The exploration is conducted differently, in a statistically natural way that softens the fluctuations of empirical means and avoids arbitrary parameters. The idea is to introduce, at time step t, a confidence region $\mathcal{CR}_{\underline{\mu}}(t)$ for $\underline{\mu}$, find the bandit problem $\underline{\tilde{\mu}}(t)$ "maximizing exploration" (in a way that we define) inside $\mathcal{CR}_{\underline{\mu}}(t)$, and track its associated optimal weight vector $\underline{\tilde{w}}(t) \stackrel{\text{def}}{=} \underline{w}(\underline{\tilde{\mu}}(t))$. Based on the lemmas of Section 3.4, which give the evolution of the weights when moving one or several arms of a bandit problem, we give a procedure that allows computing $\underline{\tilde{\mu}}(t)$. As the confidence regions shrink to $\{\underline{\mu}\}$ with time, the targeted weights will converge to $\underline{w}(\underline{\mu})$ as for the `Track-and-Stop` algorithm. A major benefit is that the procedure results in a stabilized sampling strategy, which is much easier to follow and understand (see Figure 2.7).

This stabilization, together with the careful analysis of the quantitative regularity of the solution to the optimization problem (2.20) developed in Section 3, allows us to propose a non-asymptotic analysis of `Exploration-Biased-Sampling` with finite risk bounds. A simplification of the obtained bound reads that for all standard Gaussian bandit problems $\underline{\nu}$ with means in $[0,1]$, there exist an event $\mathcal{E}$ of high probability (independent from δ) and $\delta_0 > 0$ such that algorithm `Exploration-Biased-Sampling` with the threshold of Equation (2.27) satisfies

$$\forall \delta \in (0, \delta_0], \qquad \mathbb{E}_{\underline{\mu}}\left[\tau_\delta \mathbb{I}\left\{\mathcal{E}\right\}\right] \lesssim T(\underline{\mu}) \log \tfrac{1}{\delta} + o_{\delta \to 0}(1)\ , \tag{2.28}$$

See Theorem 4.5 for a precise statement, with a closed-form expression for $o_{\delta\to 0}(1)$; note also that δ_0 depends, among others, on the probability of $\mathcal{E}$. We observe that bound (2.28) matches the asymptotic complexity. Similarly to `Track-and-Stop`, it can be shown that our strategy `Exploration-Biased-Sampling` is asymptotically optimal, although it is not a direct consequence of our non-asymptotic bound, as we considered $\mathbb{E}_{\underline{\mu}}\left[\tau_\delta \mathbb{I}\left\{\mathcal{E}\right\}\right]$ instead of $\mathbb{E}_{\underline{\mu}}[\tau_\delta]$.

In the same direction, Degenne et al. (2019) obtained a highly general and remarkable non-asymptotic bound for a pure exploration algorithm. Independently, Wang et al. (2021) obtained a sampling rule based on a Frank-Wolfe method for which they proved finite risk analysis and asymptotic optimality. Our bound has a better asymptotic behavior, but a worse behavior in the regime where gaps go to zero[13] (see Jourdan and Degenne, 2023, Table 1).

2.2.7. Towards Computationally More Efficient and More Natural Strategies

The sampling rule of the `Track-and-Stop` strategy requires computing, at each step, the solution of optimization problem (2.20). While a call to `Optimal-Weights` is not too costly experimentally (approximately proportional in the number of arms), obtaining easier sampling rules that do not need to solve the optimization problem at each step could allow us to obtain much more efficient strategies. In fact, one wants to find the most natural and simple sampling rules that ensure asymptotic optimality. In this section, we present sampling rules that go in that direction. We will still consider a general exponential model $\mathcal{D}_{\text{exp}}$ and a bandit problem $\underline{\mu}$ in $\mathcal{D}_{\text{exp}}$ with a unique optimal arm.

[13] However, we do not pretend that those bounds are comparable as we only considered $\mathbb{E}_{\underline{\mu}}\left[\tau_\delta \mathbb{I}\left\{\mathcal{E}\right\}\right]$.

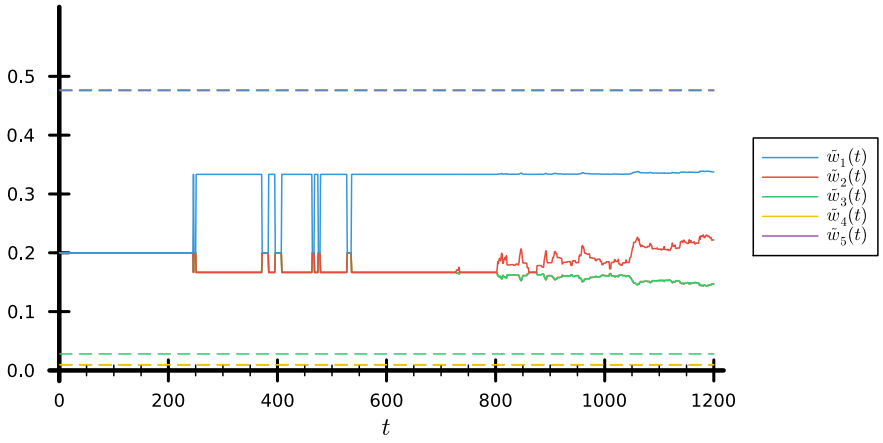


Figure 2.7: Stability of the targeted weights $\underline{\tilde{w}}(t)$ during the first 1200 rounds of a simulation of `Exploration-Biased-Sampling` with parameters $\delta = 0.01$, $\underline{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$ and D-tracking. The values of the optimal weights $\underline{w}(\underline{\mu}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ are dashed. See Figure 2.6 for a comparison with `Track-and-Stop`.

Gradient-based sampling rules. The solution of optimization problem (2.20) can be sequentially approximated using (sub-)gradient optimization methods like the Frank-Wolfe algorithm. Two such sampling strategies, `Lazy-Mirror-Ascent` and `Frank-Wolfe-based-Sampling`, were respectively proposed by Ménard (2019) and Wang et al. (2021) with improved efficiency compared to `Track-and-Stop`, and proved to be asymptotically optimal.

`top-two` **algorithms.** A promising set of strategies are `top-two` algorithms which come with very simple sampling rules: at time step t, the algorithm chooses the next arm to sample A_t between two arms, namely a *leader* L_t and a *challenger* C_t. Originally, Russo (2016, 2020) considered an algorithm for which the choice of leader and challenger is done using Thompson Sampling:

- given some prior Π_0 on the value of $a^\star(\underline{\mu})$, the leader is chosen according to the posterior Π_{t-1} computed given I_{t-1},
- the challenger is obtained similarly by drawing according to Π_{t-1} until obtaining a different arm than the leader.

The leader and challenger can also be chosen by non-Bayesian procedures (see, e.g., Jourdan et al., 2022). For the leader, we may cite the `Empirical-Best` (`EB`) and `Upper-Confidence-Bound` (`UCB`) procedures, which respectively select the arm with best empirical mean or best upper confidence bound on the mean. It might be natural to choose the challenger among arms $a \neq L_t$ minimizing the empirical transportation cost to the leader, which we call the `Transportation-Cost` (`TC`) challenger, or a version favoring exploration by adding a penalization to the cost of over-sampled arms, leading to the `Transportation-Cost-Penalized` (`TCP`) challenger. The previously mentioned leaders and challengers are gathered in Algorithms 7 and 8. The list is far from being exhaustive (see also Qin et al., 2017; Shang et al., 2020 for additional Bayesian procedures).

Remark. In the special case of best-arm identification, the `LUCBtop-two` algorithm, introduced a few years earlier by Kalyanakrishnan et al. (2012), can be seen as a Top-Two algorithm, in which the leader is pulled at even steps and the challenger is pulled at odd steps.

Algorithm 6: `top-two` sampling rule at step $t > K$

Input: history of observations I_{t-1}
`leader`, `challenger`, `sample-arm` procedures
Output: next arm to observe A_t

$L_t \leftarrow \texttt{leader}(I_{t-1})$
$C_t \leftarrow \texttt{challenger}(I_{t-1}, L_t)$
$A_t \leftarrow \texttt{sample-arm}(I_{t-1}, L_t, C_t)$

Algorithm 7: `leader` procedures for `top-two` algorithms

Input: history of observations I_{t-1}
optional: bonus function g, prior probability Π_0
Output: leader L_t

/* Empirical-Best (EB) leader */
Choose $L_t \in \underset{a \in [K]}{\operatorname{argmax}} \hat{\mu}_a(t-1)$
/* Upper-Confidence-Bound (UCB) leader */
Choose $L_t \in \underset{a \in [K]}{\operatorname{argmax}} \hat{\mu}_a(t-1) + \sqrt{\frac{g(t-1)}{N_a(t-1)}}$
/* Thompson-Sampling (TS) leader */
Choose $L_t \sim \Pi_{t-1}$

Algorithm 8: `challenger` procedures for `top-two` algorithms

Input: history of observations I_{t-1}
leader L_t
Output: challenger C_t

/* Transportation-Cost (TC) challenger */
Choose $C_t \in \underset{a \neq L_t}{\operatorname{argmin}} \mathrm{TC}_{a \to L_t}\left(\underline{\hat{\mu}}(t-1), \frac{\underline{N}(t-1)}{t-1}\right)$
/* Transportation-Cost-Penalized (TCP) challenger */
Choose
$$C_t \in \underset{a \neq L_t}{\operatorname{argmin}} \mathrm{TC}_{a \to L_t}\left(\underline{\hat{\mu}}(t-1), \frac{\underline{N}(t-1)}{t-1}\right) + \log N_a(t-1)$$
/* Thompson-Sampling (TS) challenger */
repeat
  Draw $C_t \sim \Pi_{t-1}$
until $C_t \neq L_t$

To define `top-two` strategies, it remains to choose which arm to sample between the leader and the challenger (see Algorithm 6). A simple way of doing so is to fix some probability parameter $\beta \in (0,1)$ and choose the leader with that probability[14]:

$$A_t = \begin{cases} L_t & \text{with probability } \beta\,, \\ C_t & \text{otherwise.} \end{cases} \tag{2.29}$$

We will refer to `top-two` instances using this sampling rule as `TT-leader-challenger-`β. For instance, `TT-EB-TCP-`β denotes the `top-two` algorithm used with the `Empirical-Best` leader, the `Transportation-Cost-Penalized` challenger, and sampling rule (2.29).

As a good `leader` might satisfy $L_t = a^\star$ except for a sub-linear number of time steps, using sampling rule (2.29) comes with some limitation: it implies that the best arm will be pulled a fraction β of the time, which can be far from the optimal frequency $w_{a^\star}(\underline{\mu})$. As a consequence, an easy adaption of lower bound (2.19) implies the following lower bound (see Russo, 2016) for such strategies:

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta] \geq T_\beta(\underline{\mu})\,\mathrm{kl}(\delta, 1-\delta)\,, \qquad \text{where} \quad T_\beta(\underline{\mu})^{-1} \overset{\text{def}}{=} \sup_{\substack{\underline{v}\in\Sigma_K \\ v_{a^\star}=\beta}} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a\, d(\mu_a, \lambda_a)\,, \tag{2.30}$$

and β–*asymptotically* optimal strategies are algorithms that pull the best arm with proportion β and having an upper bound asymptotically matching this lower bound. Of course, by uniqueness of $\underline{w}(\underline{\mu})$, we get $T_\beta(\underline{\mu}) = T(\underline{\mu})$ if and only if $\beta = w_{a^\star}(\underline{\mu})$. In other words, those algorithms can achieve asymptotic optimality only if $\beta = w_{a^\star}(\underline{\mu})$. An arbitrary choice of parameter β is however not too dramatic: Russo (2016) proved that for $\beta = \frac{1}{2}$, one gets:

$$T_{\frac{1}{2}}(\underline{\mu}) \leq 2T(\underline{\mu})\,,$$

hence $\frac{1}{2}$–asymptotically optimal strategies only loose a multiplicative factor 2 in theoretical performance compared to asymptotically optimal strategies.

Noticeably, to ensure β–asymptotical optimality of `TT-leader-challenger-`β algorithms, the conditions on the `leader` and `challenger` procedures are quite mild (see Jourdan et al., 2022); the combination of those proposed in Algorithms 7 and 8 are suitable.

Remark. The `TT-EB-TC-`β algorithm does not present any exploratory mechanism in the leader or challenger definitions. Even if it is β–asymptotically optimal, empirical performance at fixed values of δ suffers from some outliers at which the strategy fails to stop quickly. Strategies with an explorative procedure like `TT-EB-TCP-`β or `TT-UCB-TC-`β do not present this misbehavior and thus are preferred in practice.

Remark. Jourdan and Degenne (2023) also proved a non-asymptotic bound for Gaussian variables and the `TT-UCB-TC-`β algorithm, which asymptotically matches the β–optimality up to a factor $\frac{1}{\beta}$.

Adaptivity of `top-two` algorithms. Until recently, it has been possible to analyze `top-two` algorithms only with the use of sampling rule (2.29). It is quite frustrating that those `top-two` algorithms can only achieve β–optimality. To tackle this issue, we must rely on an adaptive sampling rule, like, for instance, adaptive pulling frequencies $\big(\beta(I_t, L_t, C_t)\big)_{t\geq 1}$ of the leader. How to design such frequencies might be driven by the targeted proportions that we want to asymptotically reach, namely $w(\underline{\mu})$. See details in Section 6.2.

Recently, You et al. (2023) proposed an adaptive `top-two` algorithm based on the original Bayesian leader and challenger of Russo (2016) called `Top-Two-Thompson-Sampling`. Their analysis proved the asymptotical optimality of the strategy for standard Gaussian variables. The proof can be adapted to obtain the same guarantees for adaptive versions of `TT-EB-TC` and `TT-EB-TCP` .

[14] One can also use a tracking procedure of proportions β instead of a sampling.

Contribution 2.14. Based on the transformation $\underline{W}$ introduced in Section 6.2.1, we propose in Section 6.2.3 a new `challenger` procedure together with adaptive pulling frequencies.

We also present partial results that generalize the analysis of You et al. (2023) in order to prove that those algorithms are asymptotically optimal for all exponential models (not only Gaussian distributions with known and common variance). See Section 6.3 for details. Yet, some work remains to complete the analysis.

2.3. Best-Arm Identification with a Fixed-Budget

We now move to the problem of best-arm identification with a fixed-budget. While the identification task seems similar to the fixed-confidence setting, it turns out that this setting is much less understood. Strategies are defined similarly to the fixed-confidence setting, except that the stopping time τ is not random and is equal to a given deterministic, known-in-advance, budget $T \geq 1$ (see Algorithm 9). The strategy returns an estimate $\hat{a}_T$ of $a^\star(\underline{\nu})$, and its quality is measured by the *probability of misidentification* or *probability of error*

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,.$$

Algorithm 9: Structure of a fixed-budget strategy

Input: budget parameter T
`sampling-rule` and `decision-rule`
Output: estimated best arm $\hat{a}_T$

for $t \in [T]$ **do**
  $A_t \leftarrow$ `sampling-rule`(I_{t-1})
  Observe $Y_t \sim \nu_{A_t}$
$\hat{a}_T \leftarrow$ `decision-rule`(I_T)

2.3.1. An Exponential Decay Rate

The simplest sampling, `Uniform-Sampling` , consists of pulling all arms equally often and returning the best empirical estimate. We will see, by using Hoeffding's inequality, that the error probability vanishes exponentially fast as T goes to $+\infty$ for a σ^2–sub-Gaussian model $\mathcal{D}_{\sigma^2}$.

Analysis of the `Uniform-Sampling` strategy. We can upper bound the probability of error when comparing arms $a^\star$ and $a \neq a^\star$ by applying Hoeffding's inequality (2.4), similarly to what we did in Equation (2.5) with the `Successive-Elimination` strategy. This gives, as all arms are pulled $\left\lfloor \frac{T}{K} \right\rfloor$ times

$$\begin{aligned}
\mathbb{P}_{\underline{\nu}}(\hat{a}_T = a) &\leq \mathbb{P}_{\underline{\nu}}\Big(\hat{\mu}_a(T) \geq \max_{b \neq a} \hat{\mu}_b(T)\Big) \\
&\leq \mathbb{P}_{\underline{\nu}}\Big(\hat{\mu}_a(T) \geq \hat{\mu}_{a^\star}(T)\Big) \\
&\leq \mathbb{P}_{\underline{\nu}}\left(\left\lfloor \frac{T}{K} \right\rfloor \Big(\hat{\mu}_a(T) - \hat{\mu}_{a^\star}(T) + \Delta_a\Big) \geq \left\lfloor \frac{T}{K} \right\rfloor \Delta_a\right) \\
&\leq \exp\left(-\frac{\Delta_a^2}{4\sigma^2}\left\lfloor \frac{T}{K} \right\rfloor\right). \qquad (2.31)
\end{aligned}$$

Applying a union bound, we obtain the following bound for the probability of misidentification:

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq (K-1) \max_{a \neq a^\star(\underline{\nu})} \mathbb{P}_{\underline{\nu}}(\hat{a}_T = a) = (K-1) \exp\left(-\frac{\Delta_{\min}^2}{4\sigma^2} \left\lfloor \frac{T}{K} \right\rfloor \right), \tag{2.32}$$

where we recall that $\Delta_{\min} = \min_{a \neq a^\star(\underline{\nu})} \Delta_a$.

Asymptotic rate of the exponential decay. We proved that the probability of misidentification of the `Uniform-Sampling` strategy goes to 0 exponentially fast with respect to the budget T. From now on, we will thus be interested in obtaining the best rate for this exponential decay, that is, in minimizing the quantity

$$\frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \,.$$

In the sequel, we will mainly work in the *asymptotic* regime where the budget T goes to $+\infty$ and hence consider *sequences* of strategies[15] indexed by their budgets T. On the one hand, we focus on obtaining instance-dependent lower bounds, valid for "reasonable" sequences of strategies, on the quantities

$$\ell(\underline{\nu}) \overset{\text{def}}{=} \liminf_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \,,$$

On the other hand, for a given sequence of strategies, we look for upper bounds of

$$u(\underline{\nu}) \overset{\text{def}}{=} \limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \,.$$

Note that $\ell(\underline{\nu}) \leq u(\underline{\nu})$ and that those rates are negative.

Example. Given Equation (2.32), we proved the upper bound on the asymptotic rate of the sequence of `Uniform-Sampling` strategies:

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{\sigma^2}, \qquad \limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{\Delta_{\min}^2}{4\sigma^2 K} \,. \tag{2.33}$$

2.3.2. `Successive-Rejects`-Type Strategies

Can we do better than `Uniform-Sampling`? Upper bound (2.33) only depends on the minimal gap $\Delta_{\min}$ and the number of arms, that is, one can change the values of the means of the $K-2$ worst arms without modifying the bound. A simple look into Equation (2.31) shows that the exponential decay of the probability of recommending sub-optimal arm a depends crucially on its gap Δ_a. In order to equalize this exponential rate among sub-optimal arms, one can be tempted to pull arm a a fraction proportional to $\frac{1}{\Delta_a^2}$ of the budget. Unfortunately, this requires good estimates of the gaps that the strategy cannot access.

This idea of having a (possibly pre-defined) different number of pulls for each arm was already studied in the fixed-confidence setting with `Successive-Elimination` strategies (see Section 2.2.1). Those strategies have also been considered in the fixed-budget setting by Audibert et al. (2010) under the name `Successive-Rejects`, as we now discuss. We continue to work with a σ^2–sub-Gaussian model $\mathcal{D}_{\sigma^2}$.

[15] In fact, we did the same in the fixed-confidence regime when considering asymptotic behaviors when δ goes to 0.

Algorithm 10: `Successive-Rejects` algorithm

Input: budget parameter T
phase lengths $(\ell_r)_{r\in[K-1]}$ such that $\sum_{r\in[K-1]} \ell_r = T$
Output: estimated best arm $\hat{a}_T$

$t \leftarrow 0$
$S_0 \leftarrow [K]$
for *each round* $r \in [K-1]$ **do**
  Observe each arm $\left\lfloor \frac{\ell_r}{K-r+1} \right\rfloor$ times
  Increase t by ℓ_r
  Choose $a_r \in \operatorname*{argmin}_{a\in S_{r-1}} \hat{\mu}_a(t)$
  $S_r \leftarrow S_{r-1} \setminus \{a_r\}$
Define $\hat{a}_T$ as the unique element of S_{K-1}

`Successive-Rejects` **strategies.** We keep the notation of the `Successive-Elimination` algorithm (see Algorithm 3). Namely, the strategy works in $K-1$ phases: at round r, candidate arms in a set S_{r-1} are pulled uniformly and the worst empirical arm of S_{r-1} (since the beginning) is dropped to obtain the set of next candidates S_r. We will consider versions of `Successive-Rejects` in which the lengths of the phases are set beforehand; they are denoted by $\ell_1, \ldots, \ell_{K-1} \geq 1$ and satisfy $\ell_1 + \ldots + \ell_{K-1} = T$. More precisely, during phase $r \in [K-1]$, the strategy draws $\lfloor \frac{\ell_r}{K-r+1} \rfloor$ times each of the $K-r+1$ arms in S_{r-1} (and does not use the few remaining time steps, if there are some). At the end of phase r, an arm of S_{r-1} has been pulled

$$n_r = \left\lfloor \frac{\ell_1}{K} \right\rfloor + \ldots + \left\lfloor \frac{\ell_r}{K-r+1} \right\rfloor$$

times since the beginning of the first phase. The description of the strategy is summarized in Algorithm 10.

Combining adapted versions of decomposition (2.6) and inequality (2.8) in the analysis of `Successive-Elimination` to σ^2–sub-Gaussian variables as in (2.31), we prove the following bound on the probability of error of `Successive-Rejects`

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star) \leq K \sum_{r\in[K-1]} \exp\left(-\frac{\Delta^2_{(K-r+1)}}{4\sigma^2} n_r\right) \leq K^2 \max_{r\in[K-1]} \exp\left(-\frac{\Delta^2_{(K-r+1)}}{4\sigma^2} n_r\right),$$

where we recall that we use the (reverse) notation of order statistics:

$$\mu_{(1)} > \mu_{(2)} \geq \mu_{(3)} \geq \cdots \geq \mu_{(K)}\,.$$

The bound can be rewritten, in terms of rates (recall that n_r depends on T),

$$\limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star) \leq -\frac{1}{4\sigma^2} \min_{r\in[K-1]} \left\{ \Delta^2_{(K-r+1)} \liminf_{T\to+\infty} \frac{n_r}{T} \right\}. \tag{2.34}$$

In order to exploit this general upper bound, one needs to carefully choose the phase lengths $\ell_1, \ldots, \ell_{K-1}$ (or equivalently the number of pulls $n_1, \ldots, n_{K-1}$).

With the knowledge of the gaps. As in the fixed-confidence setting (see page 38), we can select the phase lengths giving the best theoretical bound with the knowledge of the set of gaps. This boils down to equalizing the quantities appearing in the minimum of inequality (2.34), i.e., to take $n_r \propto \Delta_{(K-r+1)}^{-2}$. In order to not exceed the budget T, we should take

$$\forall r \in [K-1], \qquad n_r \overset{\text{def}}{=} \left\lfloor \left(\sum_{a \neq a^\star} \frac{1}{\Delta_{(a)}^2} + \frac{1}{\Delta_{(2)}^2} \right)^{-1} \frac{T}{\Delta_{(K-r+1)}^{-2}} \right\rfloor .$$

The bound (2.34) reads

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star) \leq -\frac{1}{4\sigma^2} \left(\sum_{a \neq a^\star} \frac{1}{\Delta_{(a)}^2} + \frac{1}{\Delta_{(2)}^2} \right).$$

Remark. The complexity appearing in the right-hand side was already involved in the expression (2.9) giving the value of τ_δ for the `Successive-Elimination` algorithm.

With the phase lengths of Audibert et al. (2010). How much do we lose without the knowledge of the gaps? The `Successive-Rejects` strategy was studied by Audibert et al. (2010) with the pre-defined phase lengths

$$\ell_1 \overset{\text{def}}{=} \frac{T}{\overline{\log} K}, \qquad \text{and} \qquad \forall r \in \{2, \dots, K-1\}, \qquad \ell_r \overset{\text{def}}{=} \frac{T}{(K-r+2)\,\overline{\log} K}, \tag{2.35}$$

where we define

$$\overline{\log} K \overset{\text{def}}{=} \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k},$$

which ensures that the phase lengths sum to T. In a nutshell, the first phase is the longest and is devoted to a uniform exploration of all arms, and then the $K-2$ next rounds share the rest of the budget with a slight increase in length from one phase to the next (see Figure 2.8).

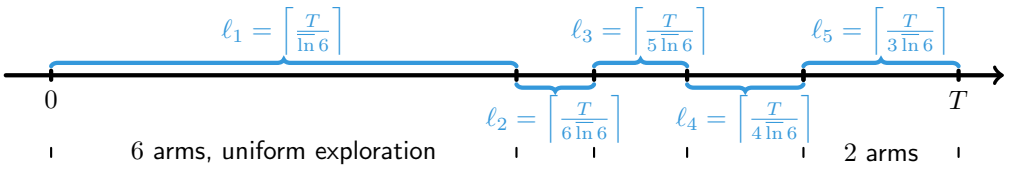


Figure 2.8: `Successive-Rejects` phase lengths of Audibert et al. (2010) for $K = 6$ arms.

Substituting the numbers of pulls $(n_r)_{r \in [K-1]}$ induced by the choice of phase lengths (2.35) into upper bound (2.34) yields the following result.

Theorem 2.15. [Audibert et al., 2010, Theorem 2]
The sequence of `Successive-Rejects` *strategies with phase lengths given by Equation* (2.35) *satisfies, for all* $\underline{\nu}$ *in* $\mathcal{D}_{\sigma^2}$ *with a unique optimal arm:*

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{4\sigma^2\,\overline{\log} K} \min_{2 \leq k \leq K} \frac{\Delta_{(k)}^2}{k}. \tag{2.36}$$

Remark.

- Audibert et al. (2010) also obtained a lower bound (that we describe in page 64) involving the quantity $\min_{2 \leq k \leq K} \frac{\Delta_{(k)}^2}{k}$. The phase lengths (2.35) are in fact chosen so as to obtain the same gap-based quantity in the upper bound, which results in the apparition of the normalizing constant $\overline{\log} K$.

- In all generality, bounds (2.33) and (2.36) are not comparable: taking a bandit problem with all gaps equal will lead to a better bound for `Uniform-Sampling` while taking one tiny gap compared to the others gives an advantage to `Successive-Rejects` (as intuition suggests). Still, bound (2.36) captures the bandit structure by considering the value of all gaps and is smaller than (2.33) in a lot of regimes. Experimentally, `Uniform-Sampling` does not outperform `Successive-Rejects` even in the first situation.
- Karnin et al. (2013) proposed `Sequential-Halving`, a variant of `Successive-Rejects` in which half of the remaining arms are dropped at each round. Its theoretical guarantees are similar to the ones for `Successive-Rejects` (up to a factor $\log 2$ in the exponential rate), while its experimental behavior shows slight improvement with a large number of arms.

Despite being the first non-trivial proposed strategy for the fixed-budget setting, the instance-dependent guarantee of `Successive-Rejects` has not been significantly improved. It seems, however, that this gap-based upper bound might be non-optimal, as we will discuss in Section 2.3.4.

Contribution. The upper bound (2.36) for the `Successive-Rejects` strategy only applies to sub-Gaussian models. We will show in Chapter 5.3 how to generalize it to other models (see Contribution 2.22 for more details). This generalization will reveal a new information-theoretic quantity replacing gaps that we present in page 67.

2.3.3. Lower Bounds

We now discuss existing lower bounds in the fixed-budget setting.

Avoiding bad strategies. Considering the class of all strategies precludes obtaining lower bounds that hold simultaneously for all bandit instances. Indeed, a strategy that always recommends a given arm (independently of all observations) will make no error for problems admitting that arm as the best, but of course, such strategies are poor, as they completely fail on any other bandit instance. To prevent this, it is possible to modify the objective and lower bound, for a given instance $\underline{\nu}$, the maximal rate of error probability among permuted instances $(\underline{\nu}^\sigma)_{\sigma\in\mathfrak{S}_K}$, where $\mathfrak{S}_K$ denotes the set of permutations of $[K]$ and $\underline{\nu}^\sigma \overset{\text{def}}{=} (\nu_{\sigma(1)},\dots,\nu_{\sigma(K)})$:

$$\max_{\sigma\in\mathfrak{S}_K} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}^\sigma}\left(\hat{a}_T \neq a^\star(\underline{\nu}^\sigma)\right).$$

This is not really satisfying in terms of writing, but sometimes handy (see Audibert et al., 2010). Yet, a more natural way of proceeding is to restrict the considered class of strategies[16].

(Exponentially) consistent sequences of strategies. Recall that, in the asymptotic point of view, we consider sequences of strategies indexed by $T \geq 1$ given a value of $K \geq 2$. We will assume that these sequences are "reasonable" in the sense below. The probability $\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))$ of misidentifying the unique optimal arm may vanish asymptotically (and even vanish exponentially fast) for all bandit problems —in not too large a model $\mathcal{D}$—, as illustrated in Section 2.3.1. We will therefore only be interested in such sequences of strategies, called (exponentially) consistent.

[16] As in the regret literature, where lower bounds are derived for uniformly fast convergent strategies (see, e.g., Burnetas and Katehakis, 1996).

Definition 2.16. [(exponentially) consistent sequence of strategies]
Fix $K \geq 2$. A sequence of strategies is *consistent*, respectively, *exponentially consistent*, on a model $\mathcal{D}$ if for all problems $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal arm,

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \underset{T\to+\infty}{\longrightarrow} 0\,, \qquad \text{respectively,} \qquad \limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) < 0\,.$$

The fundamental inequality. The fundamental inequality (2.12) that we presented in the fixed-confidence setting can also be used in fixed-budget best-arm identification to obtain lower bounds on the rate of the exponential decay. It entails the following asymptotic lemma.

Lemma 2.17. *Fix $K \geq 2$ and a model $\mathcal{D}$. Consider a consistent sequence of strategies on $\mathcal{D}$, and two bandit problems $\underline{\nu}$ and $\underline{\zeta}$ in $\mathcal{D}$ with unique and distinct optimal arms such that $a^\star(\underline{\zeta}) \neq a^\star(\underline{\nu})$. Then*

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\limsup_{T\to+\infty} \sum_{a\in[K]} \frac{\mathbb{E}_{\underline{\zeta}}[N_a(T)]}{T} \operatorname{KL}(\zeta_a, \nu_a)\,. \tag{2.37}$$

Proof. We will first prove the following non-asymptotic lower bound, which does not make any assumption on the strategy:

$$\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\frac{\sum_{a\in[K]} \mathbb{E}_{\underline{\zeta}}[N_a(T)] \operatorname{KL}(\zeta_a, \nu_a) + \log 2}{\mathbb{P}_{\underline{\zeta}}(\hat{a}_T = a^\star(\underline{\zeta}))}\,. \tag{2.38}$$

We apply the fundamental inequality (2.12) with $E = \{\hat{a}_T \neq a^\star(\underline{\nu})\}$, and by exchanging the roles of $\underline{\nu}$ and $\underline{\zeta}$:

$$\sum_{a\in[K]} \mathbb{E}_{\underline{\zeta}}[N_a(T)] \operatorname{KL}(\zeta_a, \nu_a) \geq \operatorname{kl}\Big(\mathbb{P}_{\underline{\zeta}}(\hat{a}_T \neq a^\star(\underline{\nu}))\,, \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))\Big)\,.$$

Combining this inequality with the following bound satisfied by the Kullback-Leibler divergence of the Bernoulli model[17]:

$$\forall p, q \in (0,1), \qquad \operatorname{kl}(p,q) \geq p \log \frac{1}{q} - \log 2\,, \quad \text{i.e.,} \quad \log q \geq -\frac{\operatorname{kl}(p,q) + \log 2}{p}\,, \tag{2.39}$$

and the fact that $\mathbb{P}_{\underline{\zeta}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq \mathbb{P}_{\underline{\zeta}}(\hat{a}_T = a^\star(\underline{\zeta}))$ as $a^\star(\underline{\nu}) \neq a^\star(\underline{\zeta})$, we obtain

$$\begin{aligned}
\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) &\geq -\frac{\sum_{a\in[K]} \mathbb{E}_{\underline{\zeta}}[N_a(T)] \operatorname{KL}(\zeta_a, \nu_a) + \log 2}{\mathbb{P}_{\underline{\zeta}}(\hat{a}_T \neq a^\star(\underline{\nu}))} \\
&\geq -\frac{\sum_{a\in[K]} \mathbb{E}_{\underline{\zeta}}[N_a(T)] \operatorname{KL}(\zeta_a, \nu_a) + \log 2}{\mathbb{P}_{\underline{\zeta}}(\hat{a}_T = a^\star(\underline{\zeta}))}\,,
\end{aligned}$$

which concludes the proof of (2.38).

Inequality (2.37) follows simply by considering asymptotics of (2.38)[18], and the fact that the considered sequence of strategies is consistent on $\mathcal{D}$:

$$\mathbb{P}_{\underline{\zeta}}(\hat{a}_T = a^\star(\underline{\zeta})) \underset{T\to+\infty}{\longrightarrow} 1\,, \qquad \text{and} \qquad \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \underset{T\to+\infty}{\longrightarrow} 0\,. \qquad \square$$

[17] As $\operatorname{kl}(p,q) = p \log \frac{1}{q} + \underbrace{(1-p)\log \frac{1}{1-q}}_{\geq 0} + \underbrace{p \log p + (1-p)\log(1-p)}_{\geq -\log 2}$.

[18] One can directly use the asymptotic of the kl and obtain the result without inequality (2.39).

Remark 2.18. By using inequality (2.39), we are forced to exchange the roles of $\underline{\nu}$ and $\underline{\zeta}$ when applying the fundamental inequality (2.12). This is why expectations of pulls are now under the alternative bandit problem $\underline{\zeta}$. The task of transforming those inequalities into informative lower bounds gets harder: not only do we need to find alternative bandit problems close to $\underline{\nu}$ in terms of Kullback-Leibler divergences, but we also need to simultaneously control the pull frequencies under $\underline{\zeta}$. Note also that, as a consequence, we cannot use similar techniques as in the fixed-confidence proof of lower bound (2.19), where in inequality (2.22) we replaced the unknown frequency vector $\mathbb{E}_{\underline{\nu}}[\underline{N}(\tau_\delta)]/\mathbb{E}_{\underline{\nu}}[\tau_\delta]$ by the "best" vector $\underline{v} \in \Sigma_K$, which was only possible because this frequency vector did not depend on $\underline{\zeta}$; hence, at least at first sight, we cannot conclude the existence of an optimal weight vector associated to a bandit problem $\underline{\nu}$ as in the fixed-confidence case. As we will discuss later, it does not seem possible to avoid those issues unless for models for which the Kullback-Leibler divergence is symmetric.

Lower bound of Audibert et al. (2010). Audibert et al. (2010) proposed the following lower bound for the model

$$\mathcal{B}_{[p,\,1-p]} \overset{\text{def}}{=} \{\mathrm{Ber}(x) : x \in [p, 1-p]\}$$

of Bernoulli distributions $\mathrm{Ber}(x)$ with parameters x in $[p, 1-p]$ for some $p \in (0, \frac{1}{2})$,

Theorem 2.19. [Audibert et al., 2010, Theorem 4]
Consider the model $\mathcal{B}_{[p,\,1-p]}$ for some $p \in (0, \frac{1}{2})$. For all strategies, and for all $\underline{\nu}$ in $\mathcal{B}_{[p,\,1-p]}$ with a unique optimal arm,

$$\liminf_{T\to+\infty} \frac{1}{T} \log\Big(\max_{\sigma\in\mathfrak{S}_K} \mathbb{P}_{\underline{\nu}^\sigma}\big(\hat{a}_T \neq a^\star(\underline{\nu}^\sigma)\big)\Big) \geq -\frac{5}{p(1-p)} \min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k}\,. \tag{2.40}$$

The bound involves the same quantity $\min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k}$ as in the upper bound (2.36) of `Successive-Rejects`. Yet, these two bounds do not match as there is a factor $\overline{\log} K \simeq \log K$ between both. We will compare upper and lower bounds deeply in Section 2.3.4.

The Bretagnolle-Huber technique by Kaufmann et al. (2016, Section 5.2). A specific lower bound was obtained by Kaufmann et al. (2016) for the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian distributions with common variance $\sigma^2 > 0$, based on the Bretagnolle-Huber inequality (Bretagnolle and Huber, 1979) instead of inequality (2.39):

$$\forall p, q \in (0,1), \qquad p + 1 - q \geq \frac{1}{2}\exp\big(-\mathrm{kl}(p,q)\big)\,.$$

Remark. We used inequality (2.39) in the regime $p \to 1$ and $q \to 0$. The Bretagnolle-Huber inequality should be applied in the reverse regime where $p \to 0$ and $q \to 1$, giving a similar lower bound to (2.39) except that the logarithm depends on both p and q:

$$\log\big(2(p+1-q)\big) \geq -\mathrm{kl}(p,q)\,.$$

Kaufmann et al. (2016) proved that the rate was driven by the sum of the inverse squared gaps

$$H_\Sigma(\underline{\nu}) \overset{\text{def}}{=} \sum_{a\neq a^\star(\underline{\nu})} \frac{1}{\Delta_a^2}\,. \tag{2.41}$$

More precisely, an asymptotic version of their result is the following.

Proposition 2.20. [Asymptotic version of Kaufmann et al., 2016, Theorem 16] *For all sequences of strategies and for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ with a unique optimal arm, there exists a set of alternative bandit instances $(\underline{\nu}^{(k)})_{k\neq a^\star(\underline{\nu})}$ in $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$, where each $\underline{\nu}^{(k)}$ admits k as a best arm and satisfies $H_\Sigma(\underline{\nu}^{(k)}) \leq H_\Sigma(\underline{\nu})$, and for which*

$$\liminf_{T\to+\infty} \frac{1}{T} \log \max\left\{ \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})), \max_{k\neq a^\star(\underline{\nu})} \mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T \neq k) \right\} \geq -\frac{2}{\sigma^2} H_\Sigma(\underline{\nu})^{-1} . \tag{2.42}$$

We will give a proof and a precise interpretation of this result in Section 5.6.2, but it conveys the intuition that

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \gtrsim -\frac{2}{\sigma^2} H_\Sigma(\underline{\nu})^{-1} .$$

The proposed lower bounds (2.40) and (2.42) both involve a variance term but rely on gap-based complexity terms that differ, namely, $H_\Sigma(\underline{\nu})$ and

$$H_{\max}(\underline{\nu}) = \Big(\min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k} \Big)^{-1} = \max_{2\leq k\leq K} \frac{k}{\Delta_{(k)}^2} .$$

The second quantity also appears in the upper bound (2.36), somewhat in an arbitrary way as it crucially depends on the phase lengths, and both quantities can be linked as follow

$$H_{\max}(\underline{\nu}) \leq H_\Sigma(\underline{\nu}) \leq \log(2K) H_{\max}(\underline{\nu}) ,$$

with the two inequalities being tight in all generality (there exist instances for which equalities hold).

Contribution. In Chapter 5, we will introduce new information-theoretic complexity measures that replace $H_{\max}$ and H_Σ. Those quantities will generalize the existing lower and upper bounds to more models including, e.g., non-parametric models. See Contribution 2.22 for more details.

Generalizations to other models. In order to understand if a complexity measure is a good candidate for being involved in the complexity of the fixed-budget setting, one may wonder if the proof techniques might be applied to more general models. Indeed, the lower bounds presented so far in this section are specific to Bernoulli and Gaussian.

Concerning the bound (2.42) of Kaufmann et al. (2016), a close look at the proof reveals that it heavily relies on a property even stronger than the symmetry of the Kullback-Leibler divergence for this model. In particular, generalizations beyond the Gaussian case appear to be infeasible.

Contribution. A detailed discussion on this statement might be found in Section 5.6.2, together with a proof of Proposition 2.20.

For the lower bound (2.40) of Audibert et al. (2010), however, a key inequality in their proof follows from the Kullback-Leibler – χ^2-divergence bound:

$$\forall x, y \in [p, 1-p], \qquad \mathrm{kl}(x, y) \leq \frac{(x-y)^2}{2p(1-p)} . \tag{2.43}$$

The construction may actually be generalized to models $\mathcal{D}$ with $C_{\mathcal{D}} > 0$ such that

$$\forall \nu, \nu' \text{ in } \mathcal{D}, \qquad \mathrm{KL}(\nu, \nu') \leq C_{\mathcal{D}} \big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2 . \tag{2.44}$$

This is a property that clearly holds for some exponential families: on top of the restricted Bernoulli model discussed above, where $C_{\mathcal{B}_{[p,\,1-p]}} = \frac{1}{2p(1-p)}$ by (2.43), we may cite the model $\mathcal{D}_{\sigma^2}$ of Gaussian distributions with variance σ^2, for which $C_{\mathcal{D}_{\sigma^2}} = \frac{1}{2\sigma^2}$.

A generalization of Theorem 2.19 to such models reads as follows.

Proposition 2.21. *Consider a model $\mathcal{D}$ for which property (2.44) holds with $C_{\mathcal{D}} > 0$. Then*

$$\forall \underline{\nu} \text{ in } \mathcal{D}, \qquad \liminf_{T\to+\infty} \frac{1}{T} \log\Big(\max_{\sigma\in\mathfrak{S}_K} \mathbb{P}_{\underline{\nu}^\sigma}\big(\hat{a}_T \neq a^\star(\underline{\nu}^\sigma)\big)\Big) \geq -5\, C_{\mathcal{D}} \min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k}\,.$$

2.3.4. Comparing Upper and Lower Bounds: the Challenges of the Fixed-Budget Setting

We now compare the lower bounds proposed in the previous section with the upper bound of the `Successive-Rejects` strategy. This comparison indicates important directions of research for the fixed-budget setting.

Questions. This setting is much less understood than the fixed-confidence setting.

1. We actually do not know if there exists a complexity for the fixed-budget setting: there is a $\sim \log K$ multiplicative gap between the lower and upper bounds (2.40) and (2.36). Can we obtain matching lower and upper bounds even for simple models like Gaussian or Bernoulli models? Is there an optimal strategy that asymptotically reaches the lower bounds simultaneously in all instances? We will discuss those questions in the next section.
2. The bounds so far only depend on the gaps between arms, whereas they involve the Kullback-Leibler divergence in the fixed-confidence setting, which provides a more precise quantification of the difficulty in terms of the geometry of information of the problem. Can we obtain bounds that rely on more informative quantities than gaps?
3. Most of the proposed bounds are specific to a few models (even the generalization proposed in Proposition 2.21 is limited to models with a particular structure). Can we generalize the proof techniques to more general models, like an exponential model or even a non-parametric model?

From now on, let us forget about the first point that we will consider in the next section, and see how to tackle the two other questions. It turns out that the existing literature for the fixed-budget setting offered so far a non-parametric lower bound, in the case of $K = 2$ arms.

The non-parametric bound of Kaufmann et al. (2016) for $K = 2$ arms. Namely, in a general, possibly non-parametric, model $\mathcal{D}$, Kaufmann et al. (2016, Theorem 12) stated a lower bound for all 2–armed bandit problems $\underline{\nu} = (\nu_1, \nu_2)$, and for all consistent sequence of strategies:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \inf_{\underline{\zeta}\in \mathrm{Alt}(\underline{\nu})} \max\Big\{\mathrm{KL}(\zeta_1,\nu_1),\, \mathrm{KL}(\zeta_2,\nu_2)\Big\}\,, \tag{2.45}$$

where we recall that $\mathrm{Alt}(\underline{\nu})$ is the set of alternative bandit problems[19] to $\underline{\nu}$:

$$\mathrm{Alt}(\underline{\nu}) \overset{\text{def}}{=} \Big\{\underline{\zeta} \text{ in } \mathcal{D} : a^\star(\underline{\zeta}) \neq a^\star(\underline{\nu})\Big\}\,.$$

The bound can be directly derived by taking the infimum over alternative bandit problems $\underline{\zeta}$ in lower bound (2.37):

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \inf_{\underline{\zeta}\in \mathrm{Alt}(\underline{\nu})} \limsup_{T\to+\infty} \underbrace{\sum_{a=1}^{2} \frac{\mathbb{E}_{\underline{\zeta}}[N_a(T)]}{T}\, \mathrm{KL}(\zeta_a,\nu_a)}_{\leq \max\left\{\mathrm{KL}(\zeta_1,\nu_1),\,\mathrm{KL}(\zeta_2,\nu_2)\right\}}\,.$$

[19] Already defined in the fixed-confidence setting with mean-parametrized instances, see (2.18).

New information-theoretic quantities. We have seen in lower bound (2.37) that the arguments of the involved Kullback-Leibler divergences were in reversed order compared to the fixed-confidence setting.

Remark. Recall that we have seen in Remark 2.7 that the key quantities for the non-parametric study of best-arm identification with fixed confidence are defined based on Kullback-Leibler divergences $\mathcal{K}_{\inf}^{<}$ and $\mathcal{K}_{\inf}^{>}$ with arguments in reverse order. Note also that optimal bound regret-minimization literature only depends on the $\mathcal{K}_{\inf}^{<}$ (see, e.g., Honda and Takemura, 2015; Garivier et al., 2022).

Therefore, let us introduce, for a distribution $\nu \in \mathcal{D}$ and a real number $x \in \mathbb{R}$,

$$\mathcal{L}_{\inf}^{<}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) < x\},$$

and symmetrically, by considering rather distributions ζ with expectations larger than x,

$$\mathcal{L}_{\inf}^{>}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) > x\}.$$

The $\mathcal{L}_{\inf}^{<}$ and $\mathcal{L}_{\inf}^{>}$ quantities might be key in measuring the complexity of best-arm identification under a fixed budget and be involved in generalizations of gap-bounds to general models. As a first illustration, we note that we may actually rewrite lower bound (2.45) in terms of $\mathcal{L}_{\inf}^{<}$ and $\mathcal{L}_{\inf}^{>}$ quantities[20]:

$$\inf_{\underline{\zeta}\in \mathrm{Alt}(\underline{\nu})} \max\Big\{\mathrm{KL}(\zeta_1,\nu_1),\, \mathrm{KL}(\zeta_2,\nu_2)\Big\} = \inf_{x\in[\mu_{(2)},\mu^\star]} \max\Big\{\mathcal{L}_{\inf}^{>}(x,\nu_{(2)}),\, \mathcal{L}_{\inf}^{<}(x,\nu^\star)\Big\}.$$

Contribution 2.22. In Chapter 5, we will prove upper and lower bounds based on the information-theoretic quantities $\mathcal{L}_{\inf}^{<}$ and $\mathcal{L}_{\inf}^{>}$. Those bounds will be applicable to a wide variety of models, including non-parametric models, and will imply all existing ones presented above. More precisely:

- for the upper bound, we prove that the quantity

$$\mathcal{L}(\nu',\nu) \overset{\text{def}}{=} \inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]} \Big\{\mathcal{L}_{\inf}^{>}(x,\nu') + \mathcal{L}_{\inf}^{<}(x,\nu)\Big\}, \tag{2.46}$$

  defined for ν,ν' in $\mathcal{D}$ such that $\mathrm{E}(\nu') < \mathrm{E}(\nu)$, can replace the role of gaps in the `Successive-Rejects` analysis of Audibert et al. (2010). Given a bandit problem $\underline{\nu}$ with a unique optimal arm, we may rank the arms a in non-decreasing order of $\mathcal{L}(\nu_a,\nu^\star)$, i.e., consider the permutation σ such that

$$0 = \mathcal{L}(\nu_{\sigma_1},\nu^\star) < \mathcal{L}(\nu_{\sigma_2},\nu^\star) \le \ldots \le \mathcal{L}(\nu_{\sigma_{K-1}},\nu^\star) \le \mathcal{L}(\nu_{\sigma_K},\nu^\star),$$

  and we prove that for all so-called *regular* models $\mathcal{D}$, `Successive-Rejects` guarantees that

$$\forall \underline{\nu} \text{ in } \mathcal{D}, \qquad \limsup_{T\to+\infty} \frac{1}{T}\log \mathbb{P}\big(\hat{a}_T \neq a^\star(\underline{\nu})\big) \le -\frac{1}{\overline{\log} K} \min_{2\le k\le K} \frac{\mathcal{L}(\nu_{\sigma_k},\nu^\star)}{k}. \tag{2.47}$$

  For a sub-Gaussian model, this result improves the upper bound (2.36) as it can be proved using Pinsker's inequality that $\mathcal{L}(\nu',\nu) \ge \frac{1}{4\sigma^2}\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2$.
- we also derive several lower bounds, which depend on various natural assumptions on the considered strategies. For instance, as long as the sequence of strategies asymptotically pulls the arm associated with the smallest expectation less than a fraction $\frac{1}{K}$ of the time, and

[20] See the proof of Theorem 5.14 for details about this equality, which is obtained similarly to the transportation costs calculation (2.23) of the fixed-confidence setting.

is more efficient in the identification task when one sub-optimal arm is removed from the bandit problem, then the lower bound (2.40) can be generalized to any model $\mathcal{D}$ as:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \frac{\mathcal{L}^<_{\inf}(\mu_{(k)}, \nu^\star)}{k} .$$

This lower bound is not stated in terms of infima of combinations of $\mathcal{L}^>_{\inf}$ and $\mathcal{L}^<_{\inf}$, i.e., in terms of the function $\mathcal{L}$ introduced in (2.46). While it seems challenging to obtain lower bounds involving $\mathcal{L}$, we rather obtained the following bound for so-called *monotonous* sequences of strategies, which pull more often arms with higher means:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \inf_{x\in[\mu_{(k)},\mu_{(k-1)})} \left\{ \frac{\mathcal{L}^>_{\inf}(x, \nu_{(k)})}{k-1} + \frac{\mathcal{L}^<_{\inf}(x, \nu^\star)}{k} \right\} .$$

This bound goes in the direction of the complexity appearing in upper bound (2.47), involving $\mathcal{L}$ quantities, however, the infima are taken over smaller intervals.

The bounds presented in Chapter 5 do not close the gap between lower and upper bounds: there is still (at least) a multiplicative factor $\overline{\log} K \simeq \log K$ between them. The question of reducing this gap is still open and will be discussed in the next section.

2.3.5. On Non-Matching Bounds and Minimax Results

We recall that the lower bound (2.40) and the upper bound (2.36) differ in particular by a factor proportional to $\overline{\log} K$. As a consequence, we need to understand how to fill the gap between those two bounds. Is it even possible?

The minimax lower bound of Carpentier and Locatelli (2016). Carpentier and Locatelli (2016) discuss the gap between lower bound (2.40) and upper bound (2.36) in the case of the Bernoulli model $\mathcal{B}_{[1/4,\,3/4]}$. They improve the lower bound (2.40) by a factor of $\log K$, but not simultaneously for all bandit problems $\underline{\nu}$: they obtain the improvement just for one bandit problem $\underline{\nu}$. Their lower bound result (formally stated and discussed in Section 5.6.1) can be asymptotically stated as the existence of $\underline{\nu}$ in $\mathcal{B}_{[1/4,\,3/4]}$ and of an increasing sequence of budgets $(T_n)_{n\geq 1}$ such that

$$\exists\, \underline{\nu} \text{ in } \mathcal{B}_{[1/4,\,3/4]}, \qquad \liminf_{n\to+\infty} \frac{1}{T_n} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_{T_n} \neq a^\star(\underline{\nu})) \geq - \frac{400}{\log K} H_\Sigma(\underline{\nu})^{-1} . \tag{2.48}$$

This result is different in nature from the uniform instance-dependent lower bounds considered in the previous section and in Chapter 5, which hold simultaneously for all bandit problems of a given model. Yet, it gives directions to answer the questions raised above:

- either we need to obtain new lower bounds with the $\log K$ factor so as to match upper bound (2.36),
- or this might not be possible, which would mean that lower and upper bounds cannot perfectly match for large enough models, i.e., that there does not exist an optimal strategy that reaches the best rate (among all strategies) uniformly on all instances.

Recent works have been focussing their efforts on proving that the second scenario occurs. Before discussing those results, let us explain why the fixed-budget setting seems harder than the fixed-confidence setting.

Adaptive strategies and the difficulty of the fixed-budget setting. The upper bound (2.36) was obtained for a `Successive-Rejects` strategy with pre-defined lengths phase, so, without any attempt to adapt to collected data. It could be useful to rely on those observations to choose the round lengths or design other types of strategies. If adaptivity might complicate the theoretical arguments to obtain upper bounds, one first interesting step would be to design adaptive sampling rules that perform well empirically. This task does not seem to be easy, as we expose now.

- One needs to have a high precision on the estimates of the means to confidently rely on observed data in order to adapt sampling rules. However, the misidentification event $\{\hat{a}_T \neq a^\star(\underline{\nu})\}$ that we want to control is the event that (at least) one sub-optimal empirical estimate is above the empirical estimate of the best arm:

$$\Big\{\exists a \neq a^\star(\underline{\nu}), \quad \hat{\mu}_a(T) \geq \hat{\mu}_{a^\star(\underline{\nu})}(T)\Big\}\,.$$

  This holds when there is a failure in the precision of (at least) one of the estimates: either a sub-optimal means is over-estimated, or the optimal mean is under-estimated.
- Using forced exploration (which is equivalent to beginning the process by a round of uniform exploration) does not seem possible in this setting: even a careful choice of the forced exploration strategy will lead to imprecisions in the estimates, and the strategy might not exploit those estimates correctly (see the first item). This is why Audibert et al. (2010) rather proposed another strategy called `Upper-Confidence-Bound-Exploration`, a UCB-based algorithm with an exploration bonus. To ensure good theoretical guarantees of those strategies, one requires the knowledge of a complexity term, namely $H_\Sigma(\underline{\nu})$, in order to set the scaling of the bonus (see also Gabillon et al., 2012).

Existence of optimal sequences of strategies. Komiyama et al. (2022) conjectured that decreasing the error probability under an instance $\underline{\nu}$ should increase the error probability under another instance $\underline{\zeta}$. In other words, that no strategy might perform uniformly well under all bandit instances. Given a fixed complexity function $H(\underline{\nu}) > 0$, the probability of error can be written as

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) = \exp\left(-\frac{T}{R(\underline{\nu},T)H(\underline{\nu})}\right),$$

for some corrective term $R(\underline{\nu},T) \in [0,+\infty]$. If a strategy is such that $R(\underline{\nu},T) \leq 1$, then $H(\underline{\nu})$ is a good upper bound of the complexity of $\underline{\nu}$ as the probability of error is less than

$$\exp\left(-\frac{T}{H(\underline{\nu})}\right).$$

We can extend this statement to sequences of strategies such that

$$R(\underline{\nu}) \overset{\text{def}}{=} \limsup_{T\to+\infty} R(\underline{\nu},T) \leq 1\,.$$

In addition, if $R_{\inf}(\underline{\nu}) \overset{\text{def}}{=} \inf_{\text{strategies}} R(\underline{\nu}) > 0$, then $H(\underline{\nu})$ is a good lower bound of the complexity as the probability of error is asymptotically higher than

$$\exp\left(-\frac{T}{R_{\inf}(\underline{\nu})H(\underline{\nu})}\right).$$

Can we find a "good" complexity measure H, for which

$$R_{\inf} \overset{\text{def}}{=} \inf_{\underline{\nu} \text{ in } \mathcal{D}} R_{\inf}(\underline{\nu}) > 0\,,$$

and such that there exists a uniformly optimal sequence of strategies, for which

$$R \overset{\text{def}}{=} \sup_{\underline{\nu} \text{ in } \mathcal{D}} R_{\sup}(\underline{\nu}) \leq 1 \,?$$

Komiyama et al. (2022) conjectured that such complexity does not exist for large enough models (including an exponential model). Given a complexity H, they proposed an intractable sequence of strategies that can reach the best uniform upper rate $R_{\inf}$ among all strategies. For the complexity H_Σ defined in (2.41), they also introduced an algorithm approaching this strategy by using a pre-trained neural network. It results in a strategy that performs empirically better than `Successive-Rejects` on some instances (with an improved behavior at the beginning, the uniform phase of exploration being reduced), yet without strong theoretical guarantees.

The recent results of Degenne (2023) partly confirmed the conjecture of Komiyama et al. (2022) on a theoretical aspect: for the model of standard Gaussian distributions, they consider the oracle complexity for which $R_{\inf} = 1$, which is the exact rate of the best strategy pulling arms with fixed proportions, which is equal to

$$H(\underline{\nu}) \overset{\text{def}}{=} \left(\sup_{\underline{v} \in \Sigma_K} \inf_{\underline{\zeta} \in \mathrm{Alt}(\underline{\nu})} \sum_{a \in [K]} v_a \, \mathrm{KL}(\zeta_a, \nu_a) \right)^{-1}. \tag{2.49}$$

They proved that the best rate $\inf_{\text{strategies}} R$ goes to $+\infty$ as $K \to +\infty$. In other words, for large enough K, we cannot design an algorithm that will be able to track the optimal proportions of $\underline{\nu}$ for all bandit problems $\underline{\nu}$, unlike the fixed-confidence setting. This contradicts in particular the conjecture in the conclusion of Garivier and Kaufmann (2016) according to which the complexity of the fixed-budget best-arm identification might be (2.49).

Remark. Degenne (2023) obtained this result by considering similar arguments than the minimax lower bound (2.48) of Carpentier and Locatelli (2016).

CHAPTER 3

About the Fixed-Confidence Sample Complexity Optimization Problem for Gaussian Variables

In this chapter, we study, for a Gaussian model, the solution of the sample complexity optimization problem (2.20) in best-arm identification with fixed-confidence. We present a new characterization of the solution which allows the design of a new procedure to compute the optimal weight vector and to obtain precise regularity properties concerning its dependency on the bandit instance. Materials are extracted from Section 3 (and associated appendix) of the conference paper:

A. Barrier, A. Garivier, and T. Kocák. A Non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022

Note that the lemmas of Section 3.4 contain unpublished additional statements.

Contents

3.1. Introduction

In this chapter, we consider the "multi-armed bandit" framework, a collection of $K \geq 2$ independent probability distributions $\underline{\nu} = (\nu_1, \dots, \nu_K)$ called *arms*, of unknown means $\underline{\mu} = (\mu_a)_{a\in[K]}$, and belonging to some model $\mathcal{D}$. These arms are sampled sequentially and independently: at every discrete time step $t \in \mathbb{N}^*$, an agent chooses an arm $A_t \in [K]$ based on past information and observes an independent draw Y_t from distribution ν_{A_t}. Multi-armed bandits are used as models for many situations in which one needs to find the best among a set of options, using noisy observations.

The *best-arm identification* problem consists in identifying the arm with highest mean of a bandit problem $\underline{\nu}$ (we only consider, in the rest of this section, bandit problems with a unique optimal arm):

$$\Big\{a^\star(\underline{\nu})\Big\} = \operatorname*{argmax}_{a\in[K]} \mu_a \,.$$

In the *fixed-confidence* setting (see Even-Dar et al., 2006), a confidence parameter $\delta \in (0,1)$ is given, and the objective is to design strategies that, after some random number of steps τ_δ return an estimate $\hat{a}_{\tau_\delta} \in [K]$, which is equal to the best arm $a^\star(\underline{\nu})$ with probability at least $1-\delta$. The aim is to find a strategy that minimizes the expected number of samplings $\mathbb{E}_{\underline{\nu}}[\tau_\delta]$ among all δ–correct strategies, which are strategies satisfying

$$\forall \underline{\nu} \text{ in } \mathcal{D}, \qquad \mathbb{P}_{\underline{\nu}}\big(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\nu})\big) \leq \delta \,.$$

The sample complexity optimization problem. The sample complexity of δ–correct strategies cannot be arbitrarily good. For an exponential model $\mathcal{D}_{\exp}$, for which instances $\underline{\nu}$ are characterized by their means $\underline{\mu}$, it has been proved by Garivier and Kaufmann (2016) that all strategies satisfy

$$\forall \underline{\mu} \in \mathcal{D}_{\exp}, \qquad \mathbb{E}_{\underline{\mu}}[\tau_\delta] \geq T(\underline{\mu}) \log \frac{1}{2.4\delta} \,, \tag{3.1}$$

where $T(\underline{\mu})$ is the *characteristic time* of $\underline{\mu}$, defined by the following optimization problem as

$$T(\underline{\mu})^{-1} \stackrel{\text{def}}{=} \sup_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a \, d(\mu_a, \lambda_a) \,. \tag{3.2}$$

where d denotes the mean-parameterized Kullback-Leibler divergence of the model $\mathcal{D}_{\exp}$,

$$\Sigma_K = \Big\{\underline{v} \in [0,1]^K \,:\, v_1 + \dots + v_K = 1\Big\} \qquad \text{and} \qquad \mathrm{Alt}(\underline{\mu}) = \Big\{\underline{\lambda} \text{ in } \mathcal{D}_{\exp} \,:\, a^\star(\underline{\lambda}) \neq a^\star(\underline{\mu})\Big\} \,.$$

An optimal weight vector. Garivier and Kaufmann (2016) proved that optimization problem (3.2) admits a unique maximizer $\underline{w}(\underline{\mu})$, called the *optimal weight vector*:

$$\Big\{\underline{w}(\underline{\mu})\Big\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a \, d(\mu_a, \lambda_a) \,.$$

This maximizer plays a very important role in the design of efficient fixed-confidence strategies: whatever the value of the risk δ, the more a strategy pulls arms in proportions close to $\underline{w}(\underline{\mu})$, the closest from the lower bound (3.1) its performance will be. By giving a procedure to compute the optimal weight vector when the means are known, Garivier and Kaufmann (2016) designed `Track-and-Stop`, which in a nutshell boils down to tracking the current optimal weight vector $\underline{w}(\hat{\underline{\mu}}(t))$, where $\hat{\underline{\mu}}(t)$ is the empirical estimate of $\underline{\mu}$ at time step t. `Track-and-Stop` is the first *asymptotically optimal strategy*, for which the asymptotic upper bound matches lower bound (3.1):

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\exp}, \qquad \limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu}) \,.$$

The existence of such strategies ensures the asymptotic tightness of the lower bound (3.1).

Computing the solution $\underline{w}(\underline{\mu})$. Garivier and Kaufmann (2016) proved that, with the knowledge of $\underline{\mu}$, solving the optimization problem (3.2) reduces to determining the root of a one-variable increasing function. By applying a bisection method, one may then compute $\underline{w}(\underline{\mu})$ with arbitrary precision. The procedure, referred to as `Optimal-Weights` , has consequences on the execution time of strategies that require a computation at each time step. It might be interesting to obtain more efficient methods to compute this optimal weight vector in order to speed up strategies such as `Track-and-Stop`.

Regularity and monotonicity of $\underline{w}$ and T. For a general exponential model, Garivier and Kaufmann (2016) obtained a few properties concerning the mappings $\underline{\mu} \mapsto \underline{w}(\underline{\mu})$ and $\underline{\mu} \mapsto T(\underline{\mu})$. They proved that $\underline{\mu} \mapsto \underline{w}(\underline{\mu})$, and hence $\underline{\mu} \mapsto T(\underline{\mu})$, are continuous (at problems having a unique optimal arm), that $\underline{w}(\underline{\mu})$ charges all arms:

$$\min_{a\in[K]} w_a(\underline{\mu}) > 0\,,$$

and that, if arms are ordered so that $\mu_{(1)} > \mu_{(2)} \geq \ldots \geq \mu_{(K)}$, then

$$w_{(2)}(\underline{\mu}) \geq \ldots \geq w_{(K)}(\underline{\mu})\,,$$

but there might be instances (e.g., for a Bernoulli model) for which $w_{(1)}(\underline{\mu}) < w_{(2)}(\underline{\mu})$. Apart from those basic properties, only a little is known about quantitative regularity results, or monotonicity properties of $\underline{w}(\underline{\mu})$ when, for instance, moving the mean of one of the arms. Obtaining such results would have implications in the design of more efficient and natural (in several meanings) strategies, see, e.g., Chapter 4.

The special case of Gaussian variables. Among all exponential models, the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian variables with common variance $\sigma^2 > 0$ provides a suitable setting for analysis. Its Kullback-Leibler divergence enjoys the simple closed-form expression

$$\forall \mu, \mu' \in \mathbb{R}, \qquad d(\mu, \mu') = \frac{(\mu' - \mu)^2}{2\sigma^2}\,, \tag{3.3}$$

which allows to rewrite the optimization problem (3.2) as

$$T(\underline{\mu})^{-1} = \sup_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2}\,. \tag{3.4}$$

As this $\max$–$\min$ optimization problem has a handy objective function, it might be possible to obtain new procedures for the computation of its solution $\underline{w}(\underline{\mu})$, together with monotonicity and regularity results of $\underline{w}$ and T.

Outline and contributions. This chapter is devoted to the study of the Gaussian optimization problem (3.4). For the clarity of the presentation, we only consider the model $\mathcal{D}_{\mathcal{N}_1}$ of standard Gaussian variables, but all results naturally extend to $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$.

In Section 3.2, we describe a new procedure for solving optimization problem (3.4). This results in an accelerated algorithm for its numerical resolution that we present in Section 3.3, allowing a significant speed-up for the `Track-and-Stop` algorithm in the Gaussian case. In addition, we use our procedure to deduce monotonicity properties in Section 3.4 and develop a careful analysis in Section 3.5 of the quantitative regularity of the solution to the optimization problem (3.4). Those results will be essential for defining the `Exploration-Biased-Sampling` strategy of Chapter 4 and proving its non-asymptotic guarantees.

Notation. Let, in the rest of this chapter, $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}$ be a fixed bandit parameter with a unique optimal arm. For the simplicity of the presentation, we set

$$a^\star = a^\star(\underline{\mu})\,, \qquad \underline{w} = \underline{w}(\underline{\mu})\,, \qquad \text{and} \qquad T = T(\underline{\mu})\,.$$

We recall that, for $\underline{v} \in \Sigma_K$, the quantity

$$g(\underline{\mu}, \underline{v}) \stackrel{\text{def}}{=} \inf_{\underline{\lambda} \in \mathrm{Alt}(\underline{\mu})} \sum_{a \in [K]} v_a \frac{(\mu_a - \lambda_a)^2}{2}$$

can be seen as a minimum of transportation costs of the sub-optimal arms, see Equation (2.24) and more generally Section 2.2.4, which in the Gaussian case reads

$$g(\underline{\mu}, \underline{v}) = \min_{a \neq a^\star} \underbrace{\frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \frac{\Delta_a^2}{2}}_{= \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})}\,, \tag{3.5}$$

by use of (3.3), where $\Delta_a = \Delta_a(\underline{\mu}) \stackrel{\text{def}}{=} \mu_{a^\star} - \mu_a$ is the *gap* of arm a. Thus, the optimization problem (3.4) can be equivalently written as

$$T(\underline{\mu})^{-1} = \sup_{\underline{v} \in \Sigma_K} g(\underline{\mu}, \underline{v}) = \frac{1}{2} \sup_{\underline{v} \in \Sigma_K} \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \Delta_a^2\,, \tag{3.6}$$

and the optimal weight vector $\underline{w}$ satisfies:

$$\left\{\underline{w}\right\} = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} g(\underline{\mu}, \underline{v}) = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} \frac{1}{2} \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \Delta_a^2\,. \tag{3.7}$$

Garivier and Kaufmann (2016) proved that, at the optimum $\underline{w}$, all transportation costs are equal (see also the proof of Proposition 3.2), so that, for all arms $a \neq a^\star$,

$$T(\underline{\mu})^{-1} = g(\underline{\mu}, \underline{w}) = \frac{1}{2} \frac{w_{a^\star} w_a}{w_{a^\star} + w_a} \Delta_a^2\,. \tag{3.8}$$

Remark. We might sometimes use that $g(\underline{\mu}, \underline{v})$ is defined for bandit problems $\underline{\mu}$ that admit several optimal arms (with $a^\star$ begin any of the optimal arms). In that case, $g(\underline{\mu}, \underline{v}) = 0$, all vectors $\underline{v}$ are solutions of optimization problem (3.6) and $T = +\infty$. However, unless explicitly stated, we only consider problems with a unique optimal arm.

Finally, we define the following quantities:

$$\Delta_{\min} \stackrel{\text{def}}{=} \min_{a \neq a^\star} \Delta_a\,, \qquad \Delta_{\max} \stackrel{\text{def}}{=} \max_{a \in [K]} \Delta_a\,, \qquad \text{and} \qquad w_{\min} \stackrel{\text{def}}{=} \min_{a \in [K]} w_a\,.$$

3.2. Solving the Optimization Problem

In this section, we give a new characterization of the solution $\underline{w}$ of optimization problem (3.6). We show that $\underline{w}$ can be derived using the root of the function $\phi_{\underline{\mu}}$ defined by:

$$\forall r \in \left(\frac{1}{\Delta_{\min}^2}, +\infty\right), \qquad \phi_{\underline{\mu}}(r) \stackrel{\text{def}}{=} \sum_{a \neq a^\star} \frac{1}{\left(r\Delta_a^2 - 1\right)^2} - 1\,.$$

The following properties of $\phi_{\underline{\mu}}$ are straightforward.

Lemma 3.1. *$\phi_{\underline{\mu}}$ is strictly convex, decreasing on $(\frac{1}{\Delta_{\min}^2}, +\infty)$, and thus has a unique root.*

To compute the solution $\underline{w}$ of optimization problem (3.6), it is sufficient to solve $\phi_{\underline{\mu}}(r) = 0$. The next result explains how to link r and $\underline{w}$.

Proposition 3.2. *Let $r = r(\underline{\mu})$ be the solution of $\phi_{\underline{\mu}}(r) = 0$. Then*

$$w_{a^\star} = \frac{1}{1 + \sum_{a \neq a^\star} \frac{1}{r\Delta_a^2 - 1}}, \tag{3.9}$$

$$\forall a \neq a^\star, \qquad w_a = \frac{w_{a^\star}}{r\Delta_a^2 - 1}, \tag{3.10}$$

$$\text{and} \qquad T = 2\frac{r}{w_{a^\star}}. \tag{3.11}$$

Besides,

$$w_{a^\star}^2 = \sum_{a \neq a^\star} {w_a}^2. \tag{3.12}$$

Before proving this result, we make two observations. Other important consequences of Proposition 3.2 will be derived in the next sections.

Monotonicity of the weights. In the case of $K = 2$ arms, we recall[1] that $\underline{w}(\underline{\mu}) = (0.5, 0.5)$. For $K \geq 3$, we obtain that the optimal weights are monotonous with respect to the means.

Corollary 3.3. *Assume that $K \geq 3$. Then*

$$\forall a, b \in [K], \qquad \mu_a > \mu_b \quad \Longrightarrow \quad w_a > w_b.$$

Proof. When a and b are sub-optimal, the result is a direct consequence of Equation (3.10). It remains to see that $w_{a^\star} > \max_{a \neq a^\star} w_a$ using Equation (3.12) with the fact that all weights are positive and that $K \geq 3$. □

Optimal ratio of sub-optimal arms. Equation (3.10) also implies that

$$\forall a, b \neq a^\star, \qquad \frac{w_a}{w_b} = \frac{\Delta_b^2 - \frac{1}{r}}{\Delta_a^2 - \frac{1}{r}}.$$

Intuitively, it requires about $\frac{1}{\Delta_a^2}$ samplings of arms $a^\star$ and a before being able to distinguish them, so that one could expect $\frac{w_a}{w_b}$ to be $\frac{\Delta_b^2}{\Delta_a^2}$. This would be the case if the comparisons between arms were independent. In our problem, sampling the best arm benefits the comparison with all arms, so it is worth sampling the optimal arm a little more than any single comparison would require, and hence each sub-optimal arm a little less. As a result, the ratio $\frac{w_a}{w_b}$ is closer to 1, and the factor can be seen as a "discount" on each squared gap for sharing the comparisons.

We close this section by proving Proposition 3.2.

[1] This can be obtained by Proposition 3.2, or by directly solving optimization problem (3.6).

Proof of Proposition 3.2. Let us define, for some $\beta \in [0,1]$:

$$G(\beta) \stackrel{\text{def}}{=} \max_{\substack{\underline{v} \in \Sigma_K \\ v_{a^\star} = \beta}} \min_{a \neq a^\star} \frac{\beta v_a}{\beta + v_a} \Delta_a^2 \,, \tag{3.13}$$

so that, by Equation (3.8),

$$T^{-1} = \max_{\underline{v} \in \Sigma_K} g(\underline{\mu}, \underline{v}) = \frac{1}{2} \max_{\beta \in [0,1]} G(\beta) \,. \tag{3.14}$$

Remark. $G(\beta) = 2(T_\beta)^{-1}$ where T_β is the characteristic time of strategies allowing a proportion β of their budget to the best arm, like non-adaptive `top-two` algorithms, see Equation (2.30).

By unicity of the optimal weight vector $\underline{w}$, see (3.7), we know that:

$$\left\{ w_{a^\star} \right\} = \operatorname*{argmax}_{\beta \in (0,1)} G(\beta) \,. \tag{3.15}$$

Fix $\beta \in [0,1]$. The maximum in Equation (3.13) is reached for a vector $\underline{v}$ such that $v_{a^\star} = \beta$ and all the costs $(\frac{\beta v_a}{\beta + v_a} \Delta_a^2)_{a \neq a^\star}$ are equal: $G(\beta)$ is such that

$$\forall a \neq a^\star, \qquad G(\beta) = \frac{\beta v_a}{\beta + v_a} \Delta_a^2 \,,$$

and hence

$$\forall a \neq a^\star, \qquad v_a = \frac{\beta \, G(\beta)}{\beta \Delta_a^2 - G(\beta)} \,. \tag{3.16}$$

The fact that $\underline{v} \in \Sigma_K$ yields:

$$\Phi(\beta, G(\beta)) \stackrel{\text{def}}{=} \beta + \sum_{a \neq a^\star} \frac{\beta \, G(\beta)}{\beta \Delta_a^2 - G(\beta)} - 1 = 0 \,. \tag{3.17}$$

By the implicit function theorem, there exists a mapping $\beta \in [0,1] \mapsto G(\beta)$ such that

$$\begin{aligned}
&\Phi(\beta, G(\beta)) = 0 \,, \\
\text{and} \qquad &G'(\beta) = -\frac{\frac{\partial \Phi}{\partial \beta}(\beta, G(\beta))}{\frac{\partial \Phi}{\partial G}(\beta, G(\beta))} \\
&\qquad = -\frac{1 + \sum_{a \neq a^\star} \frac{G(\beta)(\beta \Delta_a^2 - G(\beta)) - \beta \, G(\beta) \Delta_a^2}{\left(\beta \Delta_a^2 - G(\beta)\right)^2}}{\beta^2 \sum_{a \neq a^\star} \frac{\Delta_a^2}{\left(\beta \Delta_a^2 - G(\beta)\right)^2}} \\
&\qquad = -\frac{1 - \sum_{a \neq a^\star} \frac{1}{\left(\frac{\beta}{G(\beta)} \Delta_a^2 - 1\right)^2}}{\beta^2 \sum_{a \neq a^\star} \frac{\Delta_a^2}{\left(\beta \Delta_a^2 - G(\beta)\right)^2}} \,.
\end{aligned}$$

Hence $\beta \mapsto G(\beta)$ is a smooth non-negative function with a continuous derivative. By Equation (3.13), it vanishes when $\beta \to 0$ and $\beta \to 1$, and hence its maximum is reached at a point $\beta^\star$ where $G'(\beta^\star) = 0$. Defining

$$r \stackrel{\text{def}}{=} \frac{\beta^\star}{G(\beta^\star)} \,,$$

we get that r is the unique solution of $\phi_{\underline{\mu}}(r) = 0$ using the relation

$$G'(\beta^\star) = 0 \quad \iff \quad 1 - \sum_{a \neq a^\star} \frac{1}{\left(\frac{\beta^\star}{G(\beta^\star)}\Delta_a^2 - 1\right)^2} = 0\,.$$

The claimed results follow by remarking that, by Equation (3.15), $\beta^\star$ is (unique and) equal to $w_{a^\star}$. Equations (3.9), (3.10) and (3.11) can be respectively derived from (3.17), (3.16) and (3.14). Lastly, we obtain Equation (3.12) by combining Equation (3.10) and the characterization $\phi_{\underline{\mu}}(r) = 0$:

$$\sum_{a \neq a^\star} w_a^2 = w_{a^\star}^2 \sum_{a \neq a^\star} \frac{1}{\left(r\Delta_a^2 - 1\right)^2} = w_{a^\star}^2 \left(\phi_{\underline{\mu}}(r) + 1\right) = w_{a^\star}^2\,.$$

□

Remark. We largely used the explicit (and nice) expression 3.3 of the Kullback-Leibler divergence between standard Gaussian variables. Obtaining a similar proof in the general case of an exponential model might not allow us to obtain simple relations as in Proposition 3.2.

3.3. Bounds and Computation of the Problem Characteristics

By Proposition 3.2, it suffices to compute r to obtain the values of both T and $\underline{w}$. As $\phi_{\underline{\mu}}$ is a strictly convex and strictly decreasing function, Newton's iterates initialized with a value $r_0 < r$ converge to r from below at quadratic speed (the number of correct digits roughly doubles at every step). This implies that a few iterations are sufficient to guarantee machine precision. The cost of the algorithm can hence be considered proportional to that of evaluating $\phi_{\underline{\mu}}(r)$, which is linear in the number of arms. The procedure called `Gaussian-Optimal-Weights` is summarized in Algorithm 11, where we use the close form expression of the derivative of $\phi_{\underline{\mu}}$:

$$\phi'_{\underline{\mu}} : r \in \left(\frac{1}{\Delta_{\min}^2}, +\infty\right) \longmapsto -2 \sum_{a \neq a^\star} \frac{\Delta_a^2}{\left(r\Delta_a^2 - 1\right)^3}\,.$$

It remains to show that it is possible to find $r_0 < r$, and possibly close to r. The next proposition offers such a lower bound as simple functions of the gaps. This also yields tight bounds on the optimal weight vector $\underline{w}$ and the characteristic time T.

Proposition 3.4. *Let r be the solution of $\phi_{\underline{\mu}}(r) = 0$. Then the following holds:*

$$\max\left(\frac{2}{\Delta_{\min}^2}, \frac{1 + \sqrt{K-1}}{\overline{\Delta^2}}\right) \leq r \leq \frac{1 + \sqrt{K-1}}{\Delta_{\min}^2}\,, \tag{3.18}$$

$$\frac{1}{1 + \sqrt{K-1}} \leq w_{a^\star} \leq \frac{1}{2}\,, \tag{3.19}$$

and

$$\max\left(\frac{8}{\Delta_{\min}^2}, 4\frac{1 + \sqrt{K-1}}{\overline{\Delta^2}}\right) \leq T \leq 2\frac{\left(1 + \sqrt{K-1}\right)^2}{\Delta_{\min}^2}\,, \tag{3.20}$$

where $\overline{\Delta^2} \overset{\text{def}}{=} \frac{1}{K-1}\sum_{a \neq a^\star} \Delta_a^2$ is the average squared gap.

Note that all of these inequalities can be reached for certain parameters $\underline{\mu}$, as discussed after the proof of the proposition.

Algorithm 11: `Gaussian-Optimal-Weights`

Input: bandit $\underline{\mu}$ in $\mathcal{D}_{\mathcal{N}_1}$
initialization r_0
tolerance parameter tol (typically 10^{-10})
Output: optimal weight vector $\underline{w}$
characteristic time T

$r \leftarrow r_0$
while $|\phi_{\underline{\mu}}(r)| \geq \mathrm{tol}$ **do**
  $r \leftarrow r - \dfrac{\phi_{\underline{\mu}}(r)}{\phi'_{\underline{\mu}}(r)}$
$w_{a^\star} \leftarrow \left(1 + \sum_{a \neq a^\star} \dfrac{1}{r\Delta_a^2 - 1}\right)^{-1}$
for $a \neq a^\star$ **do**
  $w_a \leftarrow \dfrac{w_{a^\star}}{r\Delta_a^2 - 1}$
$T \leftarrow 2\frac{r}{w_{a^\star}}$

Proof. For all sub-optimal arms a, we set

$$q_a \stackrel{\mathrm{def}}{=} \frac{1}{r\Delta_a^2 - 1}\,.$$

By definition of r, we get

$$0 = \phi_{\underline{\mu}}(r) = \sum_{a \neq a^\star} q_a^2 - 1\,. \tag{3.21}$$

Hence the $(q_a^2)_{a \neq a^\star}$ are positive and sum to 1. This implies, in particular, that $q_a \leq 1$ for all $a \neq a^\star$, with strict inequality when $K \geq 3$.

We first prove inequalities (3.19). We recall that Equation (3.9) reads, in terms of the $(q_a)_{a \neq a^\star}$:

$$w_{a^\star} = \left(1 + \sum_{a \neq a^\star} q_a\right)^{-1}.$$

On the one hand, the upper bound of (3.19) follows from the fact that the $(q_a)_{a \neq a^\star}$ are less than or equal to 1 and relation (3.21):

$$w_{a^\star} \leq \left(1 + \sum_{a \neq a^\star} q_a^2\right)^{-1} = (1+1)^{-1} = \frac{1}{2}\,,$$

and on the other hand, we derive the lower bound of (3.19) by the Cauchy-Schwarz inequality:

$$w_{a^\star} \geq \left(1 + \sqrt{(K-1)\sum_{a \neq a^\star} q_a^2}\right)^{-1} = \frac{1}{1 + \sqrt{K-1}}\,.$$

We now prove inequalities (3.18). For the lower bound, note first that, since $q_a \leq 1$ or equivalently $r\Delta_a^2 \geq 2$ for every $a \neq a^\star$, we get

$$r \geq \frac{2}{\Delta_{\min}^2}\,. \tag{3.22}$$

Then, as $\overline{\Delta^2} = \frac{1}{K-1}\sum_{a\neq a^\star}\Delta_a^2$, the convexity of $x \mapsto \frac{1}{(rx-1)^2}$ ensures that

$$\frac{1}{K-1}\sum_{a\neq a^\star}\frac{1}{\left(\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\Delta_a^2-1\right)^2} \geq \frac{1}{\left(\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\overline{\Delta^2}-1\right)^2} = \frac{1}{K-1}\,,$$

and hence

$$\phi_{\underline{\mu}}\left(\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\right) = \sum_{a\neq a^\star}\frac{1}{\left(\frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\Delta_a^2-1\right)^2} + 1 \geq 0\,.$$

By decreasing of $\phi_{\underline{\mu}}$ (see Lemma 3.1), this entails that

$$r \geq \frac{1+\sqrt{K-1}}{\overline{\Delta^2}}\,,$$

which, together with (3.22), concludes the proof of the lower bound of (3.18). The decreasing of $\phi_{\underline{\mu}}$ is also helpful for the upper bound of (3.18): as

$$\phi_{\underline{\mu}}\left(\frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\right) = \sum_{a\neq 1}\frac{1}{\left(\frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\Delta_a^2-1\right)^2} - 1 \leq 0\,,$$

we get

$$r \leq \frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\,.$$

Finally, Equation (3.20) is derived by combining inequalities (3.18) and (3.19) with Equation (3.11). □

Tightness of the bounds. To conclude this section, we discuss the tightness of the inequalities obtained in Proposition 3.4.

- When $K = 2$, we note that lower and upper bounds match in inequalities (3.18), (3.19) and (3.20). However, the bounds do not provide any additional information as $\underline{w} = (0.5, 0.5)$ whatever the value of $\underline{\mu}$.
- In fact, equalities $r = \frac{2}{\Delta_{\min}^2}$, $w_{a^\star} = \frac{1}{2}$ and $T = \frac{8}{\Delta_{\min}^2}$ occur if and only if $K = 2$. This is because the $(q_a)_{a\neq a^\star}$ are positive and sum to 1, hence they cannot equal 1 unless $K = 2$: as we proved the associated inequalities by injecting that $q_a \leq 1$ for all $a \neq a^\star$, there is equality only when $K = 2$. The presence of additional arms increases r and T while decreases $w_{a^\star}$.
- If there is at least $K \geq 3$ arms, then the remaining equalities

  $$w_{a^\star} = \frac{1}{1+\sqrt{K-1}}\,, \quad r = \frac{1+\sqrt{K-1}}{\overline{\Delta^2}} = \frac{1+\sqrt{K-1}}{\Delta_{\min}^2}\,, \quad \text{and} \quad T = 2\frac{\left(1+\sqrt{K-1}\right)^2}{\Delta_{\min}^2}\,,$$

  are reached if and only if $\Delta_{\min} = \Delta_{\max}$, that is, if $\Delta_2 = \cdots = \Delta_K$. This might be obtained by studying the equality cases in the proof above, using the equality case of the Cauchy-Schwarz inequality for $w_{a^\star}$, the strict convexity of $x \mapsto \frac{1}{(rx-1)^2}$ and the decreasing of $\phi_{\underline{\mu}}$ for r, and finally Equation (3.20) for T. If the condition holds, note that T grows linearly with K.

3.4. Monotonicity of the max–min Problem

We now show monotonicity results for the characteristic time T and the optimal weight vector $\underline{w}$ when moving some arm(s) of $\underline{\mu}$. When $K = 2$, we have seen that $\underline{w} = (0.5, 0.5)$ whatever the bandit problem $\underline{\mu}$, and the variations of T are also known since $T = \frac{8}{\Delta^2}$ where Δ is the unique gap of $\underline{\mu}$.

In the rest of the section, we assume that $K \geq 3$. Let $\underline{\mu}'$ denote another bandit problem in $\mathcal{D}_{\mathcal{N}_1}$ sharing the same unique optimal arm $a^\star$ as $\underline{\mu}$, and define $\underline{\Delta}'$ its gap vector, $\underline{w}'$ its optimal weight vector, T' its characteristic time and r' the root of $\phi_{\underline{\mu}'}$. Those results will be useful to define the `Exploration-Biased-Sampling` strategy in Chapter 4.

3.4.1. Increasing the Mean of a Sub-Optimal Arm

The following lemma gives monotonicity properties of $\underline{w}$ and T when increasing the mean of a sub-optimal arm, which corresponds to decreasing its gap.

Lemma 3.5. *Assume that $\Delta'_b < \Delta_b$ for a fixed $b \neq a^\star$ while $\Delta'_a = \Delta_a$ for all $a \notin \{a^\star, b\}$. Then*

1. $w'_b > w_b$,
2. $w'_a < w_a$ *for all* $a \notin \{a^\star, b\}$,
3. $T' > T$,
4. *in addition:*
   - $w'_{a^\star} > w_{a^\star}$ *if arm b is (one of) the second best arm(s) of μ and μ',*
   - $w'_{a^\star} < w_{a^\star}$ *if arm b is (one of) the worst arm(s) of μ and μ'.*

The monotonicity results of the lemma are summarized in Figure 3.1. As an example, we give in Figure 3.2 the evolution of the optimal weights when modifying the value of one sub-optimal arm. This illustrates, in particular, the point 4 of the lemma: $w_{a^\star}(\underline{\mu}^x)$ decreases (respectively increases) when the moved arm is the worst arm (respectively the second best arm), which corresponds to $x \leq 2$ (respectively $x > 3$) in the figure. Otherwise, if $x \in [2, 3]$, the variation of $w_{a^\star}(\underline{\mu}^x)$ changes at some point that is marked by a dashed line on the figure, and we cannot conclude a monotonicity property in that case.

Proof.

1. Since $\Delta'_a \leq \Delta_a$ for all $a \neq a^\star$, we get

$$\phi_{\underline{\mu}'}(r) = \sum_{a \neq a^\star} \frac{1}{\left(r{\Delta'_a}^2 - 1\right)^2} + 1 > \sum_{a \neq a^\star} \frac{1}{\left(r\Delta_a^2 - 1\right)^2} + 1 = \phi_{\underline{\mu}}(r) = 0\,,$$

hence it holds that $r' > r$ by decreasing of $\phi_{\underline{\mu}'}$ (see Lemma 3.1). This implies that

$$\forall a \notin \{a^\star, b\}, \qquad \frac{1}{r'{\Delta'_a}^2 - 1} = \frac{1}{r'{\Delta_a}^2 - 1} < \frac{1}{r\Delta_a^2 - 1}\,. \tag{3.23}$$

As $K \geq 3$, such an arm a exists and hence, as $\phi_{\underline{\mu}}(r) = 0 = \phi_{\underline{\mu}'}(r')$:

$$\frac{1}{r'{\Delta'_b}^2 - 1} > \frac{1}{r\Delta_b^2 - 1}\,, \qquad \text{i.e.,} \qquad r'{\Delta'_b}^2 - 1 < r\Delta_b^2 - 1\,. \tag{3.24}$$

Combining (3.23) and (3.24) with Equation (3.10) entails that:

$$\forall a \notin \{a^\star, b\}, \qquad \frac{w'_a}{w'_b} = \frac{r'{\Delta'_b}^2 - 1}{r'{\Delta'_a}^2 - 1} < \frac{r\Delta_b^2 - 1}{r\Delta_a^2 - 1} = \frac{w_a}{w_b}\,,$$

$$\text{and} \qquad \frac{w'_{a^\star}}{w'_b} = r'{\Delta'_b}^2 - 1 < r\Delta_b^2 - 1 = \frac{w_{a^\star}}{w_b}\,.$$

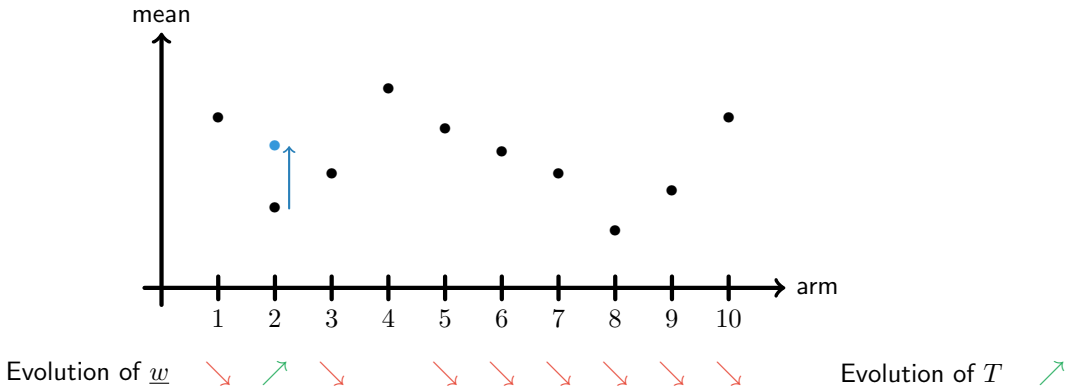


Figure 3.1: By increasing the mean of a sub-optimal arm, its associated optimal weight increases while those of other sub-optimal arms decrease, and the characteristic time increases.

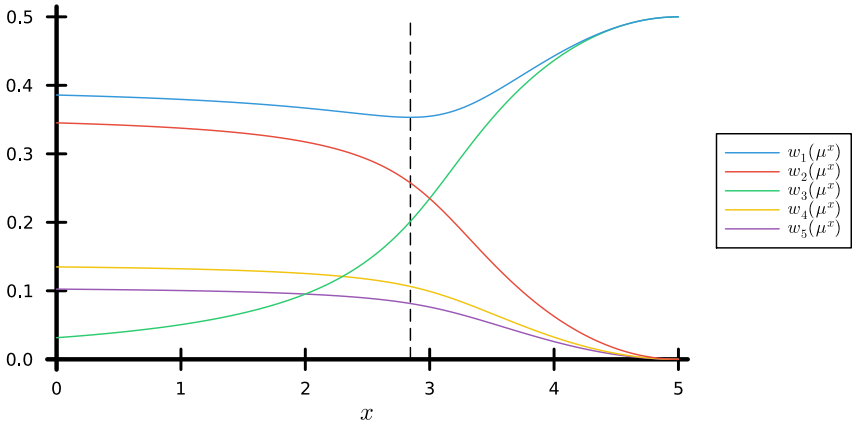


Figure 3.2: Evolution of the optimal weight $\underline{w}(\underline{\mu}^x)$ for $x \in [0,5]$, where $\underline{\mu}^x = (5, 4, x, 2.3, 2)$. The dashed line corresponds to the minimizer of $w_1(\underline{\mu}^x)$.

This leads to

$$\frac{1-w'_b}{w'_b} = \sum_{a\neq b} \frac{w'_a}{w'_b} < \sum_{a\neq b} \frac{w_a}{w_b} = \frac{1-w_b}{w_b}\,,$$

and thus, finally, to $w'_b > w_b$.

2. Given a bandit problem $\underline{\rho}$ with unique optimal arm $a^\star$, its optimal weights and the root of $\phi_{\underline{\rho}}$ can be seen as functions of its squared gap vector $\underline{d} \overset{\text{def}}{=} \underline{\Delta}(\underline{\rho})^2$: we set

$$\forall a \notin \{a^\star, b\}, \qquad F_a(\underline{d}) \overset{\text{def}}{=} \frac{1}{w_a(\underline{\rho})} = \frac{r(\underline{d})\, d_a - 1}{w_{a^\star}(\underline{\rho})} = (r(\underline{d})\, d_a - 1) + \sum_{c\neq a^\star} \frac{r(\underline{d})\, d_a - 1}{r(\underline{d})\, d_c - 1}\,.$$

where $r(\underline{d})$ is the unique solution of $\phi_{\underline{\rho}}(r) = 0$, and where the right-equalities are derived from Equations (3.9) and (3.10).

We are interested in the variations of F_a with respect to d_b (at points $\underline{d}$ corresponding to bandit problems with optimal arm $a^\star$, i.e., such that $d_{a^\star} = 0$ and $d_a > 0$ for $a \neq a^\star$). Note that, by an application of the implicit function theorem to the relation $\phi_{\underline{\rho}}(r(\underline{d})) = 0$, we know that r is differentiable with respect to d_b, with

$$\frac{\partial r}{\partial d_b}(\underline{d}) = -\frac{\dfrac{r(\underline{d})}{(r(\underline{d})\, d_b - 1)^3}}{\displaystyle\sum_{c\neq a^\star} \frac{d_c}{(r(\underline{d})\, d_c - 1)^3}} < 0\,. \tag{3.25}$$

Remark. Thanks to the first part of the proof, we already knew that r is decreasing with respect to d_b (or equivalently to the gap Δ_b).

As a consequence, the functions $(F_a)_{a\notin\{a^\star,b\}}$ are differentiable with respect to d_b, and the proof of the second point of the lemma will follow by proving that

$$\forall a \notin \{a^\star, b\}, \qquad \frac{\partial F_a}{\partial d_b}(\underline{d}) < 0\,. \tag{3.26}$$

Indeed, this will lead, as $\Delta'_b < \Delta_b$ and $\Delta'_a = \Delta_a$ for all $a \notin \{a^\star, b\}$, to

$$\forall a \notin \{a^\star, b\}, \qquad w'_a = \frac{1}{F_a(\underline{\Delta}'^2)} < \frac{1}{F_a(\underline{\Delta}^2)} = w_a\,.$$

We now prove (3.26). In the following calculations, we omit the dependency of the quantities on $\underline{d}$ to facilitate the reading, even if the dependency with respect to the squared gaps $\underline{d}$ is crucial for, e.g., $r = r(\underline{d})$. We get, for all $a \notin \{a^\star, b\}$:

$$\begin{aligned}
\frac{\partial F_a}{\partial d_b} &= \frac{\partial r}{\partial d_b} d_a + \sum_{c\neq a^\star}\left[\frac{\frac{\partial r}{\partial d_b} d_a}{rd_c - 1} - \frac{rd_a - 1}{(rd_c-1)^2}\left(\frac{\partial r}{\partial d_b} d_c\right)\right] - \frac{rd_a - 1}{(rd_b-1)^2} r \\
&= \frac{\partial r}{\partial d_b} d_a \left(1 + \sum_{c\neq a^\star} \frac{1}{rd_c - 1} - \frac{rd_c}{(rd_c-1)^2}\right) + \frac{\partial r}{\partial d_b} \sum_{c\neq a^\star} \frac{d_c}{(rd_c-1)^2} - \frac{rd_a-1}{(rd_b-1)^2} r \\
&= \frac{\partial r}{\partial d_b} d_a \sum_{c\neq a^\star} \underbrace{\frac{1 + (rd_c - 1) - rd_c}{(rd_c-1)^2}}_{=0} + \frac{\partial r}{\partial d_b} \sum_{c\neq a^\star} \frac{d_c}{(rd_c-1)^2} - \frac{rd_a-1}{(rd_b-1)^2} r \\
&= \frac{\partial r}{\partial d_b} \sum_{c\neq a^\star} \frac{d_c}{(rd_c-1)^2} - \frac{rd_a-1}{(rd_b-1)^2} r\,,
\end{aligned}$$

where, in the third equality, we used that, by definition of r,

$$1 = \sum_{c\neq a^\star} \frac{1}{(rd_c-1)^2}\,.$$

As $\frac{\partial r}{\partial d_b}$ is negative by (3.25), this implies that $\frac{\partial F_a}{\partial d_b} < 0$, i.e., (3.26) holds.

3. Using Equations (3.8) and (3.5), we get:

$$T'^{-1} = \frac{1}{2}\min_{a\neq a^\star} \frac{w'_{a^\star} w'_a}{w'_{a^\star} + w'_a} \Delta'^2_a \leq \frac{1}{2}\min_{a\neq a^\star} \frac{w'_{a^\star} w'_a}{w'_{a^\star} + w'_a} \Delta^2_a < \frac{1}{2}\min_{a\neq a^\star} \frac{w_{a^\star} w_a}{w_{a^\star} + w_a} \Delta^2_a = T^{-1}\,,$$

where the first inequality uses that $\Delta'_a \leq \Delta_b$ for all $a \neq a^\star$, and the second inequality is a consequence of the unicity of the optimal weight vector $\underline{w}$ and the fact that (as obtained in the previous points) $\underline{w} \neq \underline{w}'$.

4. As in the second point of the proof, we can look at the variations of the function

$$F_{a^\star}(\underline{d}) \stackrel{\text{def}}{=} \frac{1}{w_{a^\star}(\underline{\rho})} = 1 + \sum_{c\neq a^\star} \frac{1}{r(\underline{d})\, d_c - 1}\,,$$

where the second equality is derived from Equation (3.9).

- We want to prove that $w'_{a^\star} > w_{a^\star}$ when b is the second best arm of $\underline{\mu}$ (hence of $\underline{\mu}'$), which boils down to proving that the partial derivative $\frac{\partial F_{a^\star}}{\partial d_b}$ is positive when $d_b < \min_{a \notin \{a^\star, b\}} d_a$. Indeed, this will imply, as $\Delta'_b < \Delta_b$ and $\Delta'_a = \Delta_a$ for all $a \notin \{a^\star, b\}$, that

$$w'_{a^\star} = \frac{1}{F_{a^\star}(\underline{\Delta}'^2)} > \frac{1}{F_{a^\star}(\underline{\Delta}^2)} = w_{a^\star}\,.$$

  Computing the derivative of $F_{a^\star}$ with respect to d_b leads, by using (3.25), to

$$\begin{aligned}\frac{\partial F_{a^\star}}{\partial d_b} &= -\sum_{c \neq a^\star} \frac{\frac{\partial r}{\partial d_b} d_c}{(rd_c - 1)^2} - \frac{r}{(rd_c-1)^2} \\ &= \underbrace{\frac{r}{(rd_b-1)^2}}_{>0}\left[\frac{1}{rd_b - 1} \cdot \frac{\sum_{c\neq a^\star} \frac{d_c}{(rd_c-1)^2}}{\sum_{c\neq a^\star} \frac{d_c}{(rd_c-1)^3}} - 1\right],\end{aligned}$$

  which is positive when $d_b < \min_{a \notin \{a^\star, b\}} d_a$, as

$$(rd_b - 1) \cdot \sum_{c \neq a^\star} \frac{d_c}{(rd_c-1)^3} = \sum_{c \neq a^\star} \frac{(rd_b-1)d_c}{(rd_c-1)^3} < \sum_{c \neq a^\star} \frac{d_c}{(rd_c-1)^2}\,.$$

- To prove that $w'_{a^\star} < w_{a^\star}$ when b is the worst best arm of $\underline{\mu}'$ (hence of $\underline{\mu}$), we prove that the derivative $\frac{\partial F_{a^\star}}{\partial d_b}$ is negative as soon as $d_b > \max_{a \notin \{a^\star, b\}} d_a$. By the previous calculations, this indeed holds as

$$(rd_b - 1) \cdot \sum_{c \neq a^\star} \frac{d_c}{(rd_c-1)^3} = \sum_{c \neq a^\star} \frac{(rd_b-1)d_c}{(rd_c-1)^3} > \sum_{c \neq a^\star} \frac{d_c}{(rd_c-1)^2}\,.$$

□

3.4.2. Increasing the Mean of the Best Arm

This second lemma gives the monotonicity of $\underline{w}$ and T when increasing the mean of the best arm, which corresponds to increasing all gaps by a constant value.

Lemma 3.6. *Assume that $\Delta'_a = \Delta_a + d$ for every $a \neq a^\star$ and some $d > 0$. Then*

1. *if $b \in \operatorname{argmin}_{a \neq a^\star} \Delta_a$ is (one of) the second best arm(s) of μ, then $w'_b \leq w_b$,*
2. *$w'_{\min} \geq w_{\min}$,*
3. *$T' < T$,*
4. *we get*

$$\lim_{d\to+\infty} w'_{a^\star} = \frac{1}{1+\sqrt{K-1}}\,, \qquad \text{and} \qquad \forall a \neq a^\star, \quad \lim_{d\to+\infty} w'_a = \frac{1}{K-1+\sqrt{K-1}}\,.$$

In addition, all inequalities are strict whenever gaps $(\Delta_a)_{a\neq a^\star}$ are not all equal.

The monotonicity results of the lemma are illustrated in Figure 3.3. In Figure 3.4, we observe the evolution of the optimal weights when modifying the value of the optimal arm. It shows that we cannot give a general variation for sub-optimal arms that are not either the second or the worst arm: the weight of the third arm increases at the beginning and decreases. We observe that the weight of the optimal arm decreases and conjecture that this holds in all generality.

Before proving Lemma 3.6, we recall that the problem is scaling-invariant.

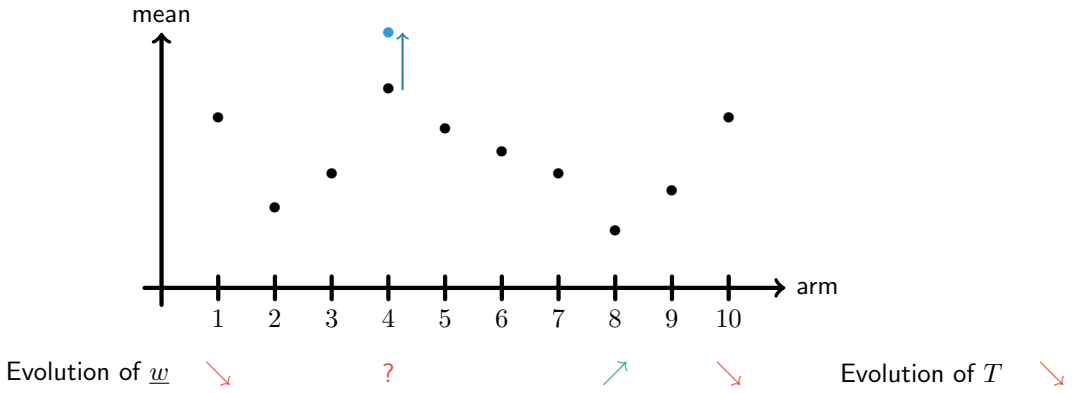


Figure 3.3: By increasing the mean of the best arm, the weight of the worst arm(s) increases, while those of the second best arm(s) decrease, and the characteristic time decreases. We also conjecture that the weight of the optimal arm decreases.

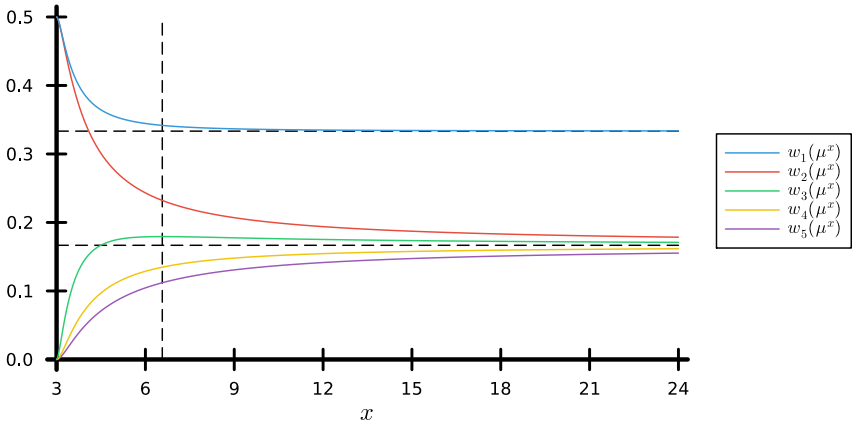


Figure 3.4: Evolution of the optimal weight $\underline{w}(\underline{\mu}^x)$ for $x \in [3, 24]$, where $\underline{\mu}^x = (x, 3, 2.7, 2.3, 2)$. The vertical dashed line corresponds to the maximizer of $w_3(\underline{\mu}^x)$, while the horizontal dashed lines are the limit values $\frac{1}{1+\sqrt{K-1}}$ and $\frac{1}{K-1+\sqrt{K-1}}$.

Lemma 3.7. *If there exists $\kappa > 0$ such that $\Delta'_a = \kappa \Delta_a$ for all $a \neq a^\star$, then $\underline{w}' = \underline{w}$.*

Proof. By definition, $\underline{w}$ and $\underline{w}'$ are solutions of the same optimization problem up to the multiplicative constant κ, as (3.7) gives:

$$\left\{\underline{w}\right\} = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} \frac{1}{2} \min_{a \neq q a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \Delta_a^2 = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} \frac{1}{2} \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} (\kappa \Delta_a)^2 = \left\{\underline{w}'\right\}.$$

□

Proof of Lemma 3.6. We first treat the easy case where all gaps $(\Delta_a)_{a \neq a^\star}$ are equal to a common gap Δ. By the equality cases of Proposition 3.4, we know that

$$w'_{a^\star} = \frac{1}{1+\sqrt{K-1}} = w_{a^\star},$$

$$\text{and} \qquad T' = \frac{2(1+\sqrt{K-1})^2}{(\Delta+d)^2} = \frac{\Delta^2}{(\Delta+d)^2} \frac{2(1+\sqrt{K-1})^2}{\Delta^2} = \frac{\Delta^2}{(\Delta+d)^2} T < T.$$

As the optimal weights of the sub-optimal arms are equal by Equation (3.10), this ensures that

$$\forall a \neq a^\star, \qquad w'_a = \frac{1}{K-1+\sqrt{K-1}} = w_a.$$

We now treat the case where at least two gaps are distinct, i.e., there exist sub-optimal arms a and b such that $\Delta_a \neq \Delta_b$.

1. Let $b \in \operatorname{argmin}_{a \neq a^\star} \Delta_a$. Let us rescale bandit problem $\underline{\mu}'$ to obtain the same gap for arm b than in $\underline{\mu}$, by multiplying the gaps of $\underline{\mu}'$ by constant $\kappa = \frac{\Delta_b}{\Delta_b + d}$. Let $\underline{\mu}''$ denote the obtained bandit (taking arbitrarily $\underline{\mu}''_{a^\star} = \underline{\mu}_{a^\star}$), and set $\underline{\Delta}'' = \underline{\Delta}(\underline{\mu}'') = \kappa \underline{\Delta}'$ and $\underline{w}'' = \underline{w}(\underline{\mu}'')$. By Lemma 3.7, we have $\underline{w}'' = \underline{w}'$. Then, on the one hand
$$\Delta''_b = \kappa \Delta'_b = \frac{\Delta_b}{\Delta_b + d}(\Delta_b + d) = \Delta_b \,,$$
and for all sub-optimal arms $a \neq b$, by non-decreasing of $x \mapsto \frac{x}{x+d}$:
$$\Delta''_a = \kappa \Delta'_a = \frac{\Delta_b}{\Delta_b + d}(\Delta_a + d) \geq \frac{\Delta_a}{\Delta_a + d}(\Delta_a + d) = \Delta_a \,,$$
with strict inequality if $\Delta_a > \Delta_b$, that is, for at least one arm a as gaps are not all equal. Applying point 2 of Lemma 3.5 to every sub-optimal arm $a \neq b$, we go from $\underline{\mu}$ to $\underline{\mu}''$ and the weight of arm b is non-decreasing, and increases during steps for which arm a is such that $\Delta''_a > \Delta_a$. Hence $w'_b = w(\underline{\mu}'') > w_b$.
2. The result can be obtained similarly to point 1 by rescaling bandit $\underline{\mu}'$ by constant $\kappa = \frac{\Delta_{\min}}{\Delta_{\min} + d}$.
3. Using Equations (3.8) and (3.5), we get:
$$T'^{-1} = \frac{1}{2} \min_{a \neq a^\star} \frac{w'_{a^\star} w'_a}{w'_{a^\star} + w'_a} {\Delta'_a}^2 > \frac{1}{2} \min_{a \neq a^\star} \frac{w_{a^\star} w_a}{w_{a^\star} + w_a} {\Delta'_a}^2 > \frac{1}{2} \min_{a \neq a^\star} \frac{w_{a^\star} w_a}{w_{a^\star} + w_a} \Delta_a^2 = T^{-1} \,,$$
where the first inequality is a consequence of the unicity of the optimal weight vector $\underline{w}$ and the fact that (as obtained in the previous points) $\underline{w} \neq \underline{w}'$, and the second uses that $\Delta'_a > \Delta_a$ for all $a \neq a^\star$.
4. Using the rescaling argument of the second point of this proof, we get that the limit, when d goes to ∞, of bandit $\underline{\mu}''$ is the bandit problem $\underline{\mu}_{\lim}$ with best arm $a^\star$ and constant gap $\Delta = \Delta_b$. By continuity of the optimal weight vector, this implies that
$$\forall a \in [K], \qquad \lim_{d \to +\infty} w'_a = w'_a(\underline{\mu}_{\lim}) \,.$$
Injecting the values of the optimal weight vector for a bandit problem with constant gaps, recalled at the beginning of this proof, gives the result. □

3.4.3. Increasing the Mean of the Worst Arms

Lastly, we look at the monotonicity of $\underline{w}$ and T when increasing the worst means.

Lemma 3.8. *Let $B = \operatorname{argmin}_{a \in [K]} \mu_a$ (respectively $B' = \operatorname{argmin}_{a \in [K]} \mu'_a$) be the set of the worst arms of $\underline{\mu}$ (respectively $\underline{\mu}'$) and assume that $B \subset B'$ and $\Delta'_{\max} < \Delta_{\max}$, while $\Delta'_a = \Delta_a$ for all $a \notin B'$.*

1. *$w'_{a^\star} \leq w_{a^\star}$,*
2. *if c is a sub-optimal arm such that $\Delta'_c = \Delta_c$, then $w'_{a^\star} \leq w_{a^\star}$,*
3. *$w'_{\min} \geq w_{\min}$,*
4. *$T' > T$,*
5. *we get*

$$\lim_{\Delta'_{\max} \to \Delta_{\min}} w'_{a^\star} = \frac{1}{1 + \sqrt{K-1}} \,, \quad \text{and} \quad \forall a \neq a^\star, \quad \lim_{\Delta'_{\max} \to \Delta_{\min}} w'_a = \frac{1}{K - 1 + \sqrt{K-1}} \,.$$

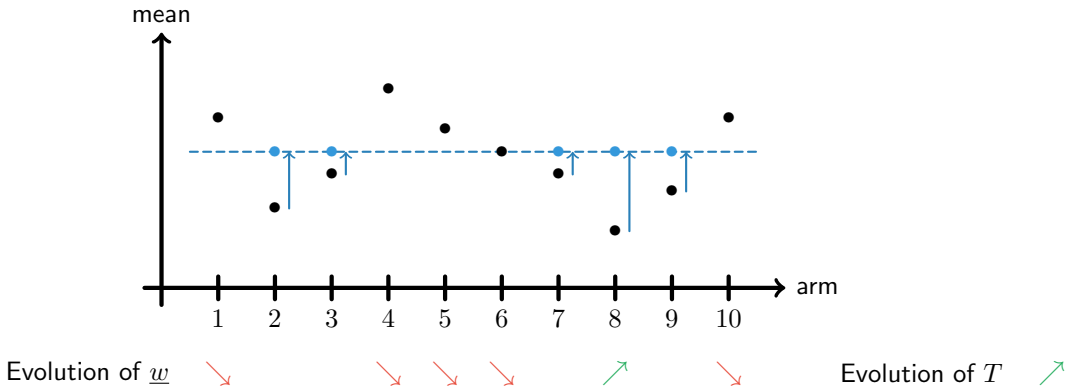


Figure 3.5: By increasing the means of the worst arms to a common value, the minimal optimal weight and the characteristic time increase, while the optimal weights of unmoved arms decrease.

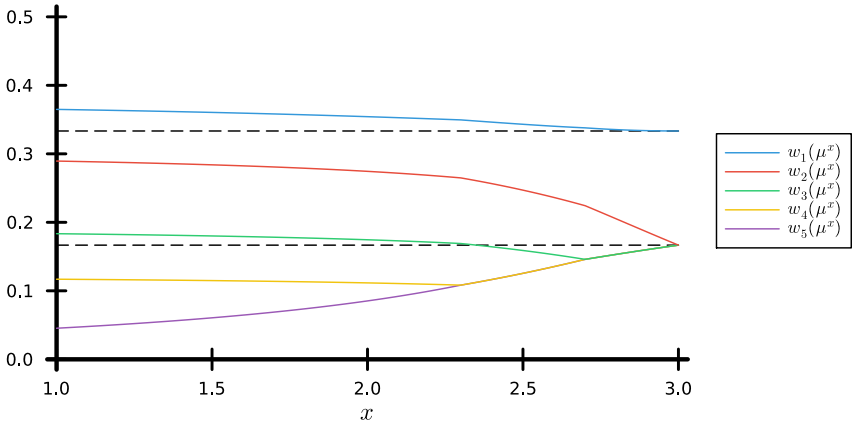


Figure 3.6: Evolution of the optimal weight vector $\underline{w}(\underline{\mu}^x)$ for $x \in [1,3]$, where $\underline{\mu}^x = (5, 3, \max(2.7, x), \max(2.3, x), x)$. The dashed lines are the limit values $\frac{1}{1+\sqrt{K-1}}$ and $\frac{1}{K-1+\sqrt{K-1}}$.

The monotonicity results of the lemma are illustrated in Figure 3.5. In Figure 3.6, we observe the evolution of the optimal weights when modifying the minimal mean of the arms. It shows in particular the limit behaviors and the fact that while the mean of an arm is not modified, its weight decreases.

Proof. One can mimic the proof of Lemma 3.5 to prove that, if we equally increase the means of all the worst arms of $\underline{\mu}$ to a value which is smaller or equal to the second worst mean of $\underline{\mu}$, then:

1. $w_{\min}$ increases,
2. w_a decreases for all $a \notin B \cup \{a^\star\}$,
3. $w_{a^\star}$ decreases,
4. T increases.

By applying this observation a finite number of steps (one needs to split the reasoning each time a new arm is added to the list of worst arms), we deduce all but the last point.

Limit behaviors of point 5 might be obtained similarly to Lemma 3.6. □

3.5. Regularity Properties

In this section, we provide quantitative regularity results for the functions $\underline{w}$, T and g. The results presented in this section will prove to be essential to the non-asymptotic analysis of the `Exploration-Biased-Sampling` strategy in Chapter 4.

3.5.1. Regularity of $\underline{w}$ and T

In this section, we show explicit bounds on the regularity of $\underline{\rho} \mapsto \underline{w}(\underline{\rho})$ and $\underline{\rho} \mapsto T(\underline{\rho})$. We keep the notation of the last section.

Theorem 3.9. *Assume that, for some* $\varepsilon \in [0, 1/7]$,

$$\forall a \neq a^\star, \qquad (1-\varepsilon)\Delta_a^2 \leq {\Delta'_a}^2 \leq (1+\varepsilon)\Delta_a^2 \,. \tag{3.27}$$

Then

$$\forall a \in [K], \qquad (1-10\varepsilon)w_a \leq w'_a \leq (1+10\varepsilon)w_a \,, \tag{3.28}$$

$$\text{and} \qquad (1-3\varepsilon)T \leq T' \leq (1+6\varepsilon)T \,. \tag{3.29}$$

Proof. All along the proof, we will use the following inequalities, that might be easily checked,

$$\forall x \in \mathbb{R}, \qquad \tfrac{1}{1+x} \leq 1-x \,, \tag{3.30}$$

$$\forall u \in [0, 1/2], \qquad \tfrac{1}{1-u} \leq 1+2u \,. \tag{3.31}$$

We will first prove the two following supporting results:

$$\frac{r}{1+\varepsilon} \leq r' \leq \frac{r}{1-\varepsilon} \,, \tag{3.32}$$

$$\text{and} \qquad \frac{1-3\varepsilon}{w_{a^\star}} \leq \frac{1-3\varepsilon}{1-\varepsilon}\frac{1}{w_{a^\star}} \leq \frac{1}{w'_{a^\star}} \leq \frac{1+5\varepsilon}{1+\varepsilon}\frac{1}{w_{a^\star}} \leq \frac{1+5\varepsilon}{w_{a^\star}} \,. \tag{3.33}$$

On the one hand, we get, using the right-inequality in (3.27),

$$\phi_{\underline{\mu}'}\left(\frac{r}{1+\varepsilon}\right) = \sum_{a\neq a^\star} \frac{1}{\left(\frac{r}{1+\varepsilon}{\Delta'_a}^2 - 1\right)^2} - 1 \geq \sum_{a\neq a^\star} \frac{1}{\left(\frac{r}{1+\varepsilon}\Delta_a^2(1+\varepsilon) - 1\right)^2} - 1 = \phi_{\underline{\mu}}(r) = 0 \,,$$

and on the other hand, using the left inequality,

$$\phi_{\underline{\mu}'}\left(\frac{r}{1-\varepsilon}\right) = \sum_{a\neq a^\star} \frac{1}{\left(\frac{r}{1-\varepsilon}{\Delta'_a}^2 - 1\right)^2} - 1 \leq \sum_{a\neq a^\star} \frac{1}{\left(\frac{r}{1-\varepsilon}\Delta_a^2(1-\varepsilon) - 1\right)^2} - 1 = \phi_{\underline{\mu}}(r) = 0 \,.$$

By decreasing of $\phi_{\underline{\mu}'}$ (Lemma 3.1) and definition of r', we deduce that (3.32) holds.

We move to the proof of (3.33). For all $a \neq a^\star$, we get

$$r'{\Delta'_a}^2 \leq \frac{1+\varepsilon}{1-\varepsilon} r\Delta_a^2 = (1+\eta) r\Delta_a^2 \,, \qquad \text{where} \qquad \eta \overset{\text{def}}{=} \frac{1+\varepsilon}{1-\varepsilon} - 1 = \frac{2\varepsilon}{1-\varepsilon} \,,$$

hence, by inequality (3.30),

$$\begin{aligned} \frac{1}{r'{\Delta'_a}^2 - 1} &\geq \frac{1}{(1+\eta)r{\Delta_a}^2 - 1} = \frac{1}{(r\Delta_a^2 - 1)\left(1 + \frac{\eta r\Delta_a^2}{r\Delta_a^2 - 1}\right)} \\ &\geq \frac{1}{r\Delta_a^2 - 1}\left(1 - \frac{\eta r\Delta_a^2}{r\Delta_a^2 - 1}\right) = \frac{1}{r\Delta_a^2 - 1} - \eta\frac{1}{r\Delta_a^2 - 1} - \eta\frac{1}{\left(r\Delta_a^2 - 1\right)^2} \,. \end{aligned} \tag{3.34}$$

By Equation (3.9), we get

$$\begin{aligned}\frac{1}{w'_{a^\star}} &= 1 + \sum_{a\neq a^\star} \frac{1}{r'{\Delta'_a}^2 - 1} \\ &\geq 1 + (1-\eta)\sum_{a\neq a^\star}\frac{1}{r\Delta_a^2-1} - \eta \underbrace{\sum_{a\neq a^\star}\frac{1}{\left(r\Delta_a^2-1\right)^2}}_{=\phi_{\underline{\mu}}(r)+1=1} \\ &= (1-\eta)\frac{1}{w_{a^\star}} = \frac{1-3\varepsilon}{1-\varepsilon}\frac{1}{w_{a^\star}}\,,\end{aligned}$$

which gives the second inequality of (3.33).

We move to the proof of the third inequality. Let $a \neq a^\star$. We get

$$r'{\Delta'_a}^2 \geq \frac{1-\varepsilon}{1+\varepsilon} r\Delta_a^2 = (1-\eta) r\Delta_a^2\,, \qquad \text{where} \qquad \eta \overset{\text{def}}{=} 1 - \frac{1-\varepsilon}{1+\varepsilon} = \frac{2\varepsilon}{1+\varepsilon}\,.$$

Note that $\eta \leq 1/4$ as $\varepsilon \leq 1/7$. Using that $r \geq \frac{2}{\Delta_{\min}}$ by inequality (3.18), we get $r\Delta_a^2 \geq 2$, hence, by decreasing of $x \mapsto \frac{x}{x-1}$ on $(2,+\infty)$,

$$\eta\frac{r\Delta_a^2}{r\Delta_a^2-1} \leq \frac{1}{2}\,.$$

By using inequality (3.31), we have

$$\begin{aligned}\frac{1}{r'{\Delta'_a}^2-1} \leq \frac{1}{(1-\eta)r\Delta_a^2-1} &= \frac{1}{\left(r\Delta_a^2-1\right)\left(1-\frac{\eta r\Delta_a^2}{r\Delta_a^2-1}\right)} \\ &\leq \frac{1}{r\Delta_a^2-1}\left(1+2\frac{\eta r\Delta_a^2}{r\Delta_a^2-1}\right) = \frac{1}{r\Delta_a^2-1} + 2\eta\frac{1}{r\Delta_a^2-1} + 2\eta\frac{1}{\left(r\Delta_a^2-1\right)^2}\,.\end{aligned} \tag{3.35}$$

Consequently,

$$\begin{aligned}\frac{1}{w'_{a^\star}} &= 1 + \sum_{a\neq a^\star}\frac{1}{r'{\Delta'_a}^2-1} \\ &\leq 1 + (1+2\eta)\sum_{a\neq a^\star}\frac{1}{r\Delta_a^2-1} + 2\eta\underbrace{\sum_{a\neq a^\star}\frac{1}{\left(r\Delta_a^2-1\right)^2}}_{\phi_{\underline{\mu}}(r)+1=1} \\ &= (1+2\eta)\frac{1}{w_{a^\star}} = \frac{1+5\varepsilon}{1+\varepsilon}\frac{1}{w_{a^\star}}\,,\end{aligned}$$

which concludes the proof of (3.33).

We now deduce the results from inequalities (3.32) and (3.33). In all the following equations, we use at least one of the inequalities of (3.32) and (3.33), together with various bounds satisfied by $\varepsilon \leq \frac{1}{7}$. First, taking the inverse of (3.33), and using (3.30) for the left-inequality and (3.31) for the right-inequality, we get

$$(1-5\varepsilon)w_{a^\star} \leq \frac{w_{a^\star}}{1+5\varepsilon} \leq w'_{a^\star} \leq \frac{w_{a^\star}}{1-3\varepsilon} \leq (1+6\varepsilon)w_{a^\star}\,.$$

Equation (3.10), with respectively (3.34) and (3.35), yield for all $a \neq a^\star$:

$$w'_a = \frac{w'_{a^\star}}{r'{\Delta'_a}^2 - 1} \geq \frac{\frac{w_{a^\star}}{1+5\varepsilon}}{(r\Delta_a^2 - 1)\Big(1 + \frac{2\varepsilon}{1-\varepsilon}\underbrace{\frac{r\Delta_a^2}{\Delta_a^2 - 1}}_{\leq 2}\Big)} = \frac{1-\varepsilon}{(1+5\varepsilon)(1+3\varepsilon)} w_a \geq (1-10\varepsilon) w_a \,,$$

and $$w'_a = \frac{w'_{a^\star}}{r'{\Delta'_a}^2 - 1} \leq \frac{\frac{w_{a^\star}}{1-3\varepsilon}}{(r\Delta_a^2 - 1)\Big(1 - \frac{2\varepsilon}{1+\varepsilon}\underbrace{\frac{r\Delta_a^2}{\Delta_a^2 - 1}}_{\leq 2}\Big)} = \frac{1+\varepsilon}{(1-3\varepsilon)^2} w_a \leq (1+10\varepsilon) w_a \,.$$

This concludes the proof of (3.28). Finally, we obtain (3.29) by using Equation (3.11), as, on the one hand,

$$T' = \frac{2r'}{w'_{a^\star}} \geq 2 \cdot \frac{r}{1+\varepsilon} \cdot \frac{1-3\varepsilon}{1-\varepsilon} \frac{1}{w_{a^\star}} = \frac{1-3\varepsilon}{1-\varepsilon^2} \cdot \frac{2r}{w_{a^\star}} \geq (1-3\varepsilon) T \,,$$

and, on the other hand,

$$T' = \frac{2r'}{w'_{a^\star}} \leq 2 \cdot \frac{r}{1-\varepsilon} \cdot \frac{1+5\varepsilon}{1+\varepsilon} \frac{1}{w_{a^\star}} = \frac{1+5\varepsilon}{1-\varepsilon^2} \cdot \frac{2r}{w_{a^\star}} \leq (1+6\varepsilon) T \,.$$

□

3.5.2. Regularity of g

The regularity of g might also be studied. Controlling the variations of g requires getting an upper bound on the gap values. For simplicity, we work here with the model $\mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ of standard Gaussian variables with means in $[0,1]$, but one might generalize it to all sub-models of $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of standard Gaussian bandit problems with bounded gaps. We get the following property:

Proposition 3.10. *Let $\underline{\mu}$ and $\underline{\mu}'$ be two bandit problems in $\mathcal{D}_{\mathcal{N}_1}^{[0,1]}$, with potentially several best arms, and let $\underline{v} \in \Sigma_K$. Setting*

$$\varepsilon = \max_{a \in [K]} \big|\mu_a - \mu'_a\big| \,, \qquad \text{and} \qquad \eta = \max_{a \in [K]} \frac{|w_a(\underline{\mu}) - v_a|}{w_a(\underline{\mu})} \,,$$

we get

$$g(\underline{\mu}', \underline{v}) \geq \frac{(1-\eta)^2}{1+\eta} \Big(g\Big(\underline{\mu}, \underline{w}(\underline{\mu})\Big) - \frac{\varepsilon}{2} \Big) \,.$$

We will prove Proposition 3.10 by combining two Lemmas. The first lemma states that $g(\,\cdot\,, \underline{u})$ is $\frac{1}{2}$–Lipschitz for the infinity norm.

Lemma 3.11. *Let $\underline{\mu}$ and $\underline{\mu}'$ be two bandit problems in $\mathcal{D}_{\mathcal{N}_1}^{[0,1]}$, with potentially several best arms. Then, for all optimal weight vectors $\underline{u} \in \Sigma_K$,*

$$\Big| g(\underline{\mu}', \underline{u}) - g(\underline{\mu}, \underline{u}) \Big| \leq \frac{1}{2} \max_{a \in [K]} \big|\mu_a - \mu'_a\big| \,.$$

Proof. As $\underline{\mu}$ and $\underline{\mu}'$ play similar roles, we only prove that

$$g(\underline{\mu}', \underline{u}) \geq g(\underline{\mu}, \underline{u}) - \frac{\varepsilon}{2} \,, \qquad \text{where} \qquad \varepsilon = \max_{a \in [K]} \big|\mu_a - \mu'_a\big| \,.$$

- Assume first that $\underline{\mu}$ and $\underline{\mu}'$ have a common best arm denoted by $a^\star$. Then by Equation (3.5),

$$\begin{aligned} g(\underline{\mu}', \underline{u}) - g(\underline{\mu}, \underline{u}) &= \frac{1}{2} \min_{a \neq a^\star} \frac{u_{a^\star} u_a}{u_{a^\star} + u_a} {\Delta'_a}^2 - \frac{1}{2} \min_{b \neq a^\star} \frac{u_{a^\star} u_b}{u_{a^\star} + u_b} \Delta_b^2 \\ &= \frac{1}{2} \min_{a \neq a^\star} \max_{b \neq a^\star} \frac{u_{a^\star} u_a}{u_{a^\star} + u_a} {\Delta'_a}^2 - \frac{u_{a^\star} u_b}{u_{a^\star} + u_b} \Delta_b^2 \\ &\geq \frac{1}{2} \min_{a \neq a^\star} \frac{u_{a^\star} u_a}{u_{a^\star} + u_a} \left({\Delta'_a}^2 - \Delta_a^2 \right). \end{aligned} \tag{3.36}$$

  by taking $b = a$. Fix $a \neq a^\star$. One has:

$$|\Delta_a - \Delta'_a| = |(\mu_{a^\star} - \mu'_{a^\star}) - (\mu_a - \mu'_a)| \leq |\mu_{a^\star} - \mu'_{a^\star}| + |\mu_a - \mu'_a| \leq 2\varepsilon \,,$$

  from which we obtain, using that the gaps are bounded by 1

$$\left| \Delta_a^2 - {\Delta'_a}^2 \right| = |\Delta_a - \Delta'_a| (\Delta_a + \Delta'_a) \leq 4\varepsilon \,.$$

  As $\underline{u}$ is an optimal weight vector, it satisfies $u_a \leq u_{a^\star} \leq \frac{1}{2}$ by Corollary 3.3 and bound (3.19), so that:

$$\frac{u_{a^\star} u_a}{u_{a^\star} + u_a} \leq \frac{1}{2} \frac{u_a}{u_{a^\star} + u_a} \leq \frac{1}{2} \frac{u_a}{2u_a} = \frac{1}{4} \,.$$

  Finally,

$$\frac{u_{a^\star} u_a}{u_{a^\star} + u_a} \left({\Delta'_a}^2 - \Delta_a^2 \right) \geq -\varepsilon \,.$$

  This concludes the proof by injecting this inequality into (3.36).

- In case $\underline{\mu}$ and $\underline{\mu}'$ do not have a common best arm, define the family of bandits $\left(\underline{\mu}^{(t)}\right)_{t \in [0,1]}$ by

$$\forall t \in [0,1], \forall a \in [K], \qquad \mu_a^{(t)} = (1-t)\mu_a + t\mu'_a \,.$$

  One can check that $\underline{\mu} = \underline{\mu}^{(0)}$, $\underline{\mu}' = \underline{\mu}^{(1)}$ and

$$\forall t_1, t_2 \in [0,1], \qquad \max_{a \in [K]} |\mu_a^{(t_1)} - \mu_a^{(t_2)}| \leq |t_1 - t_2| \cdot \varepsilon \leq \varepsilon \,. \tag{3.37}$$

  Select the subdivision $0 = t_0 < t_1 < \cdots < t_N = 1$ of times at which the optimal arms of $\underline{\mu}^{(t)}$ are modified. Note that $N \geq 2$ as $\underline{\mu}$ and $\underline{\mu}'$ do not have a common best arm. Note that by continuity:
  - for any $1 \leq n \leq N-1$, $\underline{\mu}^{(t_n)}$ has at least two best arms so that $g(\underline{\mu}^{(t_n)}, \underline{u}) = 0$,
  - $\underline{\mu}^{(1)}$ and $\underline{\mu}$ have a common best arm,
  - $\underline{\mu}^{(N-1)}$ and $\underline{\mu}'$ have a common best arm.

  Thus, applying the first part of the proof,

$$\begin{aligned} g(\underline{\mu}', \underline{u}) - g(\underline{\mu}, \underline{u}) &= \underbrace{g(\underline{\mu}', \underline{u}) - \overbrace{g(\underline{\mu}^{(t_{N-1})}, \underline{u})}^{=0}}_{\geq 0} + \overbrace{g(\underline{\mu}^{(t_1)}, \underline{u})}^{=0} - g(\underline{\mu}, \underline{u}) \\ &\geq g(\underline{\mu}^{(t_1)}, \underline{u}) - g(\underline{\mu}, \underline{u}) \\ &\geq -\frac{1}{2} \max_{a \in [K]} |\mu_a^{(t_1)} - \mu_a| \\ &\geq -\frac{\varepsilon}{2} \,, \end{aligned}$$

  where we used (3.37) in the last inequality. This concludes the proof. □

The second lemma controls the variations of $g(\underline{\mu}', \cdot)$. Note that this lemma does not require the use of bounded gaps.

Lemma 3.12. *Let $\underline{\mu}'$ be a bandit problem in $\mathcal{D}_{\mathcal{N}_1}$ with potentially several arms. Let $\underline{u}, \underline{v} \in \Sigma_K$ be such that, for a fixed $\eta \in [0,1]$,*

$$\max_{a \in [K]} \frac{|u_a - v_a|}{u_a} \leq \eta \,. \tag{3.38}$$

Then:

$$g(\underline{\mu}', \underline{v}) \geq \frac{(1-\eta)^2}{1+\eta} g(\underline{\mu}', \underline{u}) \,.$$

Proof. Let $a^\star$ is (one of) the best arm(s) of $\underline{\mu}'$. Note that condition (3.38) can be written as

$$\forall a \in [K], \qquad (1-\eta)u_a \leq v_a \leq (1+\eta)u_a \,.$$

Hence, for all $a \neq a^\star$,

$$\frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \geq \frac{(1-\eta)^2}{(1+\eta)} \cdot \frac{u_1 u_a}{u_1 + u_a} \,,$$

which entails

$$g(\underline{\mu}', \underline{v}) = \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} {\Delta'_a}^2 \geq \frac{(1-\eta)^2}{1+\eta} \min_{a \neq a^\star} \frac{u_1 u_a}{u_1 + u_a} {\Delta'_a}^2 = \frac{(1-\eta)^2}{1+\eta} g(\underline{\mu}', \underline{u}) \,.$$

□

Proof of Proposition 3.10. The result follows directly by Lemmas 3.12 and 3.11 with $\underline{u} = \underline{w}(\underline{\mu})$:

$$g(\underline{\mu}', \underline{v}) \geq \frac{(1-\eta)^2}{1+\eta} g(\underline{\mu}', \underline{w}(\underline{\mu})) \geq \frac{(1-\eta)^2}{1+\eta} \left(g(\underline{\mu}, \underline{w}(\underline{\mu})) - \frac{\varepsilon}{2} \right) .$$

□

3.6. Conclusion

In this chapter, we studied the sample complexity optimization 3.2 for a Gaussian model. The fact that this optimization problem only depends on gaps allowed us to obtain a new characterization of the solution $\underline{w}(\underline{\mu})$. This characterization implies new general bounds for $\underline{w}(\underline{\mu})$ and $T(\underline{\mu})$, together with monotonicity properties that will be useful for the definition of the strategy `Exploration-Biased-Sampling` in Chapter 4.

We also investigated the regularity of the solution, proving quantitative results that complement the continuity result of Garivier and Kaufmann (2016). Obtaining such results in the context of a general exponential model could be an interesting perspective for future work, but requires dealing with a less explicit and easy-to-handle Kullback-Leibler divergence.

CHAPTER 4

A Fixed-Confidence Strategy with Non-Asymptotic Guarantees

In this chapter, we propose a new strategy for the problem of best-arm identification with fixed-confidence of Gaussian variables. This strategy, called `Exploration-Biased-Sampling`, is not only asymptotically optimal: it is to the best of our knowledge the first strategy with non-asymptotic bounds that asymptotically matches the sample complexity. Its sampling rule is built in order to naturally encourage exploration and uses the results of Chapter 3. The content of the present chapter is extracted from Sections 1, 2 and 4 (and associated appendices) of the conference paper

A. Barrier, A. Garivier, and T. Kocák. A Non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022

Contents

4.1. Introduction

Many modern systems of automatic decisions (from recommender systems to clinical trials, through auto-ML and parameter tuning) require finding the best among a set of options, using noisy observations obtained by successive calls to a random mechanism (see, e.g., Lattimore and Szepesvári, 2020). The simplest formal model for such situations is the *standard Gaussian multi-armed bandit*, a collection of $K \geq 2$ independent Gaussian distributions called *arms* of unknown means $\underline{\mu} = (\mu_a)_{a\in[K]} \in \mathbb{R}^K$ and variances all equal to 1. They are sampled sequentially and independently: at every discrete time step $t \in \mathbb{N}^*$, a learner chooses an arm $A_t \in [K]$ based on past information, and observes an independent draw Y_t from distribution $\mathcal{N}(\mu_{A_t}, 1)$.

Best-arm identification. The *best-arm identification* problem consists in identifying the arm with highest mean of $\underline{\mu}$. Unless otherwise specified, we only consider bandit problems with a unique optimal arm:

$$\left\{a^\star(\underline{\mu})\right\} \stackrel{\text{def}}{=} \operatorname*{argmax}_{a\in[K]} \mu_a \,.$$

The corresponding distribution mean will be denoted by $\mu^\star$. In the *fixed-confidence* setting (see Even-Dar et al., 2006; Kalyanakrishnan et al., 2012), a confidence parameter $\delta \in (0,1)$ is given, and the objective is to design strategies that will stop after some (random) finite number of observations and give an estimate of the best arm $a^\star(\underline{\mu})$ which is correct with probability at least $1-\delta$. Such strategies are called δ–correct, and their performance is measured by how quick they are to take a decision.

Formally, a strategy is defined by

- a *sampling rule*, which consists in choosing the arm $A_t \in [K]$ to observe at each time step $t \geq 1$. This arm A_t depends on the previous observations $Y_1, \dots, Y_{t-1}$, but also possibly on some external randomization that we capture by the random variable U_{t-1}. A_t is thus $\mathcal{F}_{t-1}$–measurable, where $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(I_{t-1})$ with $I_{t-1} \stackrel{\text{def}}{=} (U_0, Y_1, U_1, Y_2, U_2, \dots, Y_{t-1}, U_{t-1})$. I_{t-1} corresponds to the information available at the end of the time step $t-1$.
- a *stopping rule* τ_δ, which is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$,
- a *decision rule* $\hat{a}_{\tau_\delta}$ which is $\mathcal{F}_{\tau_\delta}$–measurable.

The general structure of a strategy is presented in Algorithm 12. A strategy is *δ–correct* if for all bandit problems $\underline{\mu}$ in the model $\mathcal{D}_{\mathcal{N}_1}$ of standard Gaussian variables,

$$\mathbb{P}_{\underline{\mu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\mu})) \leq \delta \,,$$

where $\mathbb{P}_{\underline{\mu}}$ is the probability distribution under bandit problem $\underline{\mu}$.

Lower bound. The aim of the problem is to find strategies that minimize the *sample complexity*, that is, the expected number of samplings $\mathbb{E}_{\underline{\mu}}[\tau_\delta]$. The sample complexity of δ–correct strategies cannot be arbitrarily good: it has been proved by Garivier and Kaufmann (2016) that they obey the lower bound

$$\forall \underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}, \qquad \mathbb{E}_{\underline{\mu}}[\tau_\delta] \geq T(\underline{\mu}) \log \frac{1}{2.4\delta} \,, \tag{4.1}$$

where the *characteristic time* $T(\underline{\mu})$ is the solution of the following optimization problem

$$T(\underline{\mu})^{-1} = \sup_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a \frac{(\mu_a - \lambda_a)^2}{2} \,, \tag{4.2}$$

where

$$\Sigma_K = \left\{\underline{v} \in [0,1]^K \,:\, v_1 + \dots + v_K = 1\right\} \qquad \text{and} \qquad \mathrm{Alt}(\underline{\mu}) = \left\{\underline{\lambda} \text{ in } \mathcal{D}_{\mathcal{N}_1}^{[0,1]} \,:\, a^\star(\underline{\lambda}) \neq a^\star(\underline{\mu})\right\}.$$

Algorithm 12: General structure of a fixed-confidence strategy

Input: confidence parameter δ
`sampling-rule`, `stopping-condition`, `decision-rule`
Output: stopping time τ_δ
estimated best arm $\hat{a}_{\tau_\delta}$

Observe each arm once // initialization
$t \leftarrow K$
while `stopping-condition`(I_t, δ) *is not satisfied* **do**
  Increase t by 1
  $A_t \leftarrow$ `sampling-rule`(I_{t-1})
  Observe $Y_t \sim \mathcal{N}(\mu_{A_t}, 1)$
$\tau_\delta \leftarrow t$
$\hat{a}_{\tau_\delta} \leftarrow$ `decision-rule`(I_{τ_δ})

An optimal weight vector. The information-theoretic analysis of Garivier and Kaufmann (2016) also highlights the nature of the optimal sampling strategy: whatever the value of the risk δ, one should sample the arms with frequencies proportional to $\underline{v} = \underline{w}(\underline{\mu})$, the (unique and well-defined) maximizer in the right-hand side of Equation (4.2). This observation allowed the authors to introduce `Track-and-Stop`, the first asymptotically optimal strategy, which satisfies,

$$\forall \underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}, \qquad \limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} = T(\underline{\mu}),$$

proving, by passing, that the lower bound (4.1) is tight. The algorithm works as follows: at every time step t, the optimal frequencies relative to an estimate $\hat{\underline{\mu}}(t)$ of the mean parameter $\underline{\mu}$ are computed and used to determine which action is to be selected next: we pick the action that lays the most behind its estimated optimal frequency unless one action was severely under-sampled (in which case its exploration is forced). Some improvements were proposed: for example, Ménard (2019) proved that it is not necessary to solve the optimization problem in every time step.

The shortcomings of `Track-and-Stop`. The `Track-and-Stop` algorithm is not only a theoretical contribution, but it also proved to be numerically efficient, far exceeding its competitors in a wide variety of settings. It was improved in different directions (Degenne and Koolen, 2019; Degenne et al., 2019; Shang et al., 2020), and also provides a simple template for extensions, for bandit problems with structure (Kocák and Garivier, 2020), as long as the optimization problem (4.2) can be solved. Yet, `Track-and-Stop` suffers from certain shortcomings. First, a close look into the proofs shows that the theoretical guarantees proved so far are really asymptotic in nature. Second, the forced exploration appears very arbitrary, with a rate of $\sqrt{t}$ that has no other justification than lying somewhere between constant and linear functions. Third, the sampling strategy appears to be pretty unstable, especially at the beginning: the target frequencies can vary significantly as the estimated means fluctuate before stabilizing around their expectations. Fourth, `Track-and-Stop` does not present the intuitively desirable behavior to sample uniformly in the beginning, until sufficient information has been gathered for significant differences between the arms to emerge. This is in contrast with strategies like Racing (Kaufmann and Kalyanakrishnan, 2013), which are sub-optimal but intuitively appealing. Altogether, these issues lead for example to unpredictable and irregular conduct at the beginning of multiple A/B testing cases with many arms very close to optimal.

Towards non-asymptotic bounds. While the proven optimality of `Track-and-Stop` is purely asymptotic, a different approach is followed in (Karnin et al., 2013; Jamieson et al., 2014; Chen et al., 2017) for moderate values of δ. The proposed strategies are sub-optimal by a multiplicative constant but are proven to satisfy explicit non-asymptotic bounds. More recently, Degenne et al. (2019) obtained a general non-asymptotic bound, a remarkable but hardly comparable result in particular settings.

Outline and contributions. In this chapter, we try to make a link between both approaches by introducing a new strategy, `Exploration-Biased-Sampling` , with non-asymptotic guarantees which, in the regime $\delta \to 0$, correspond to the lower bound on the sample complexity. Additionally, `Exploration-Biased-Sampling` solves the issues of `Track-and-Stop` mentioned above. We work with the model $\mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ of standard Gaussian variables with means in $[0,1]$, but all results can be generalized to[1] the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}^{\mathcal{M}}$ of Gaussian variables with means in a bounded set $\mathcal{M}$ and common variance σ^2.

We present in Section 4.2 our new `Exploration-Biased-Sampling` strategy. The exploration is conducted very differently, in a statistically natural way that softens the fluctuations of empirical means and avoids arbitrary parameters. It results in a stabilized sampling strategy, that is much easier to follow and understand. Its theoretical properties and guarantees are stated in Section 4.3, including a non-asymptotic analysis with finite risk bounds for which the proof is presented in Section 4.5. These results have required developing a careful analysis of the quantitative regularity of the solution to the optimization problem (4.2) that was presented in Chapter 3. Lastly, we illustrate the performance and behavior of our strategy by numerical experiments in Section 4.4.

Note that, independently, Wang et al. (2021) obtained a sampling rule based on a Frank-Wolfe method for which they proved finite risk analysis and asymptotic optimality. Our finite risk bound has a better asymptotic behavior, but a worse behavior in the regime where gaps go to zero.

Notation. For the simplicity of the presentation, when there is no confusion we set

$$a^\star = a^\star(\underline{\mu})\,, \qquad \Delta_a = \Delta_a(\underline{\mu})\,, \qquad \underline{w} = \underline{w}(\underline{\mu})\,, \qquad \text{and} \qquad T = T(\underline{\mu})\,,$$

where $\Delta_a(\underline{\mu}) = \mu^\star - \mu_a$ is the gap of arm a. We recall (see Section 3.1) that the optimization problem rewrites:

$$T(\underline{\mu})^{-1} = \sup_{\underline{v} \in \Sigma_K} g(\underline{\mu}, \underline{v})\,, \qquad \text{where} \qquad g(\underline{\mu}, \underline{v}) = \underbrace{\frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \frac{\Delta_a^2}{2}}_{= \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})}\,,$$

and hence the optimal weight vector $\underline{w}$ satisfies:

$$\left\{\underline{w}\right\} = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} g(\underline{\mu}, \underline{v}) = \operatorname*{argmax}_{\underline{v} \in \Sigma_K} \frac{1}{2} \min_{a \neq a^\star} \frac{v_{a^\star} v_a}{v_{a^\star} + v_a} \Delta_a^2\,.$$

Garivier and Kaufmann (2016) proved that, at the optimum $\underline{w}$, all transportation costs are equal, so that, for all arms $a \neq a^\star$,

$$T(\underline{\mu})^{-1} = g(\underline{\mu}, \underline{w}) = \frac{1}{2} \frac{w_{a^\star} w_a}{w_{a^\star} + w_a} \Delta_a^2\,. \tag{4.3}$$

Remark. Unless explicitly stated, we only consider instances with a unique optimal arm. However, we might sometimes use that $g(\underline{\mu}, \underline{v})$ is also defined for bandit problems $\underline{\mu}$ that admit several optimal

[1] The algorithms discussed here can be used with a sub-Gaussian model with a known upper bound on the variances. However, for such models, the sample complexity bounds proved in this chapter apply but are not necessarily optimal.

arms (with $a^\star$ being any of the optimal arms). In that case, $g(\underline{\mu}, \underline{v}) = 0$, $T(\underline{\mu}) = +\infty$, and we define the optimal weight vector $\underline{w}(\underline{\mu})$, somewhat arbitrarily, by

$$\underline{w}_a(\underline{\mu}) = \frac{\mathbb{I}\left\{a \in \mathcal{A}^\star(\underline{\mu})\right\}}{\operatorname{card} \mathcal{A}^\star(\underline{\mu})},$$

where $\mathcal{A}^\star(\underline{\mu}) = \operatorname{argmax}_{a\in[K]} \mu_a$ is the set of optimal arms of $\underline{\mu}$.

For a given strategy facing a bandit problem $\underline{\mu}$, let $N_a(t)$ and $\hat{\mu}_a(t)$ denote the number of pulls and the empirical mean[2] of arm a at step t:

$$N_a(t) \overset{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = a\} \qquad \text{and} \qquad \hat{\mu}_a(t) \overset{\text{def}}{=} \frac{1}{N_a(t)} \sum_{s\in[t]} Y_s \mathbb{I}\{A_s = a\}\,.$$

Without loss of generality (see the paragraph on optional skipping page 37), we assume that the observation at time step t is $Y_t = X_{A_t, N_{A_t}(t)}$, where $(X_{a,n})_{a\in[K], n\geq 1}$ are independent random variables such that $X_{a,n} \sim \mathcal{N}(\mu_{A_t}, 1)$ for all $a \in [K]$ and $n \geq 1$. As a consequence, we notably get

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{n=1}^{N_a(t)} X_{a,n} \overset{\text{def}}{=} \hat{\mu}_{a,N_a(t)}\,. \tag{4.4}$$

4.2. The Exploration-Biased-Sampling Strategy

In this section, we introduce our new strategy called Exploration-Biased-Sampling. Instead of Track-and-Stop's greedy choice of actions based on a plug-in estimate of $\underline{\mu}$, it relies on a specific estimator that is biased toward uniform exploration. We fix $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ with a unique optimal arm and define the following quantities, for which the dependency with respect to $\underline{\mu}$ is omitted if there is no confusion:

$$\Delta_{\min} \overset{\text{def}}{=} \min_{a\neq a^\star} \Delta_a > 0\,, \qquad \Delta_{\max} \overset{\text{def}}{=} \max_{a\in[K]} \Delta_a\,, \qquad \text{and} \qquad w_{\min} \overset{\text{def}}{=} \min_{a\in[K]} w_a\,.$$

We recall that $w_{\min}$ is the optimal weight value of the worst arm(s) of $\underline{\mu}$.

4.2.1. Conservative Tracking

The main idea of the algorithm is to design a sampling policy of arms that naturally encourages exploration without forcing it like Track-and-Stop does. To do so, the objective is to "wrap" the optimal weight vector $\underline{w}(\underline{\mu})$ "from above", by ensuring that we never under-estimate its minimal value. Indeed, even an arm with a low mean needs to be sampled sufficiently often until one is very confident that it is sub-optimal. The idea is to construct a confidence region $\mathcal{CR}_{\underline{\mu}} \subset [0,1]^K$ for $\underline{\mu}$ on which one can efficiently find a bandit $\underline{\tilde{\mu}} \in \mathcal{CR}_{\underline{\mu}}$ maximizing the minimal weight $w_{\min}$:

$$\underline{\tilde{\mu}} \in \operatorname*{argmax}_{\underline{\rho}\in\mathcal{CR}_{\underline{\mu}}} w_{\min}(\underline{\rho})\,. \tag{4.5}$$

As long as $\underline{\mu}$ belongs to the confidence region $\mathcal{CR}_{\underline{\mu}}$, choosing the target weights $\underline{w}(\underline{\tilde{\mu}})$ guarantees that every arm is explored sufficiently, as $w_{\min}(\underline{\tilde{\mu}}) \geq w_{\min}(\underline{\mu})$. The exploration bias decreases with the number of observations, as $\mathcal{CR}_{\underline{\mu}}$ shrinks to $\{\underline{\mu}\}$, and in the end arms are sampled with frequencies close to the optimal weight vector $\underline{w}(\underline{\mu})$.

[2] As strategies initially observe each arm once, $\hat{\mu}_a(t)$ is well-defined for $t \geq K$.

This approach to exploration requires two ingredients:

- the exploration-biased bandit $\underline{\tilde{\mu}}$ needs to be efficiently computable. It turns out to be the case if the confidence region is a product of confidence intervals on each arm (a mild requirement since the arms are independent). We propose Algorithm 13, an efficient procedure for computing $\underline{\tilde{\mu}}$. Intuitively, maximizing $w_{\min}$ over $\mathcal{CR}(\underline{\mu})$ requires increasing and equalizing all the positive gaps as much as possible. The associated bandit will indeed be the one for which it is harder to identify the second-best arm and thus it will require to sample the worst arms more frequently. This gives a candidate bandit for each potential best arm, and our algorithm compares those candidates. Figure 4.1 illustrates on an example the principle of Algorithm 13, whose correctness is proved in Proposition 4.1. The algorithm requires `Gaussian-Optimal-Weights` (Algorithm 11 of Chapter 3), an efficient procedure for solving optimization problem (4.2).
- the regularity of the mapping $\underline{\rho} \mapsto \underline{w}(\underline{\rho})$ needs to be explicitly known. Indeed, the confidence region will decrease with the number of observations, and $\underline{\tilde{\mu}}$ will come close to $\underline{\mu}$. The continuity proved by Garivier and Kaufmann (2016) for the asymptotic optimality of `Track-and-Stop` is not sufficient: the first quantitative bounds were given in Section 3.5.

One can remark that as long as the confidence intervals have a non-empty intersection, which means the observations do not permit to exclude that any of them is optimal, the exploration-biased weights returned by Algorithm 13 are uniform and the arms are sampled in a round-robin way (as in a Racing or Successive Elimination algorithm like in Even-Dar et al., 2006).

Proposition 4.1. *Let $\mathcal{CR} = \prod_{a\in[K]}[\mu_a^-, \mu_a^+] \subset [0,1]^K$ be a confidence region and*

$$(\underline{\tilde{\mu}}, \underline{\tilde{w}}) \leftarrow \texttt{Exploration-Biased-Weights}(\mathcal{CR})\,.$$

Then $\underline{\tilde{w}} = \underline{w}(\underline{\tilde{\mu}})$ and $\underline{\tilde{\mu}}$ satisfies Equation (4.5).

The proof relies on the results of Section 3.4.

Proof. We assume that $K \geq 3$, otherwise $\underline{w}(\underline{\rho}) = (\frac{1}{2}, \frac{1}{2})$ for all $\underline{\rho}$ and the result is clear.

With the notation of Algorithm 13, we first observe that $\underline{\tilde{w}} = \underline{w}(\underline{\tilde{\mu}})$. When minUB $\geq$ maxLB the algorithm returns a constant bandit and $\underline{\tilde{w}} = (\frac{1}{K}, \dots, \frac{1}{K})$ is its optimal weight vector by convention. As all optimal weight vectors $\underline{w}(\underline{\rho})$ are such that $w_{\min}(\underline{\rho}) \leq \frac{1}{K}$, we obtain that $\underline{\tilde{\mu}}$ satisfies (4.5).

Now assume that minUB $<$ maxLB, i.e., at least two confidence intervals are disjoint, and fix $\underline{\rho} \in \mathcal{CR}$. If $\underline{\rho}$ as several optimal arms, then $w_{\min}(\underline{\rho}) = 0$ so that trivially $w_{\min}(\underline{\rho}) \leq w_{\min}(\underline{\tilde{\mu}})$. Assume now that $\underline{\rho}$ has a unique optimal arm denoted by a. Note that $a \in$ PotentialBest, so that we will show that

$$w_{\min}(\underline{\rho}) \leq w_{\min}(\underline{\tilde{\mu}}^{\mathsf{test}(a)}) \leq \tilde{w}_{\min}\,.$$

The latest inequality stems from the choice of $\underline{\tilde{\mu}}$ by the procedure into the set of potential most exploring bandits $\{\underline{\tilde{\mu}}^{\mathsf{test}(a)} : a \in \text{PotentialBest}\}$. To obtain the first inequality, we will transform the means of $\underline{\rho}$ to those of $\underline{\tilde{\mu}}^{\mathsf{test}(a)}$ by modifications that only increase $w_{\min}$, thanks to the results of Section 3.4. We recall that the value of $w_{\min}$ is the optimal weight value of any of the worst arms. The procedure, illustrated in Figure 4.2, is the following:

1. Transform $\underline{\rho}$ into $\underline{\rho}^{(1)}$ by increasing arm a so that $\underline{\rho}^{(1)}_a = \mu_a^+$. Using Lemma 3.6, one has
$$w_{\min}(\underline{\rho}^{(1)}) \geq w_{\min}(\underline{\rho})\,.$$
2. Transform $\underline{\rho}^{(1)}$ into $\underline{\rho}^{(2)}$ by decreasing, for each arm $b \neq a$, μ_b to $\max(\mu_b^-, \rho_{\min})$, where $\rho_{\min} = \min_{c\in[K]} \rho_c$. By several applications[3] of Lemma 3.5, one has
$$w_{\min}(\underline{\rho}^{(2)}) \geq w_{\min}(\underline{\rho}^{(1)})\,.$$

[3] Note that, as the modified means do not go below $\underline{\rho}_{\min}$, the worst arms of $\underline{\rho}$ stay in the set of worst arms.

Algorithm 13: `Exploration-Biased-Weights`

Input: confidence region $\mathcal{CR} = \prod_{a\in[K]}[\mu_a^-, \mu_a^+]$
Output: exploration-biased bandit $\underline{\tilde{\mu}} \in \mathcal{CR}$
exploration-biased optimal weight vector $\underline{\tilde{w}} = \underline{w}(\underline{\tilde{\mu}})$

maxLB $\leftarrow \max_{a\in[K]} \mu_a^-$
minUB $\leftarrow \min_{a\in[K]} \mu_a^+$
if *minUB* $\geq$ *maxLB* **then**
  $\underline{\tilde{\mu}} \leftarrow (\text{minUB}, \dots, \text{minUB})$
  $\underline{w} \leftarrow (\frac{1}{K}, \dots, \frac{1}{K})$
else
  PotentialBest $\leftarrow \{a \in [K] : \mu_a^+ > \text{maxLB}\}$
  $\underline{\tilde{w}} \leftarrow (0, \dots, 0)$
  for $a \in$ *PotentialBest* **do**
    $\tilde{\mu}_a^{\text{test}(a)} \leftarrow \mu_a^+$
    for $b \in [K] \setminus \{a\}$ **do**
      $\tilde{\mu}_b^{\text{test}(a)} \leftarrow \max(\mu_b^-, \text{minUB})$
    $\underline{w}^{\text{test}(a)} \leftarrow$ `Gaussian-Optimal-Weights`$\left(\underline{\tilde{\mu}}^{\text{test}(a)}\right)$
    if $\min_{b\in[K]} w_b^{\text{test}(a)} > \min_{b\in[K]} \tilde{w}_b$ **then**
      $\underline{\tilde{w}} \leftarrow \underline{w}^{\text{test}(a)}$
      $\underline{\tilde{\mu}} \leftarrow \underline{\tilde{\mu}}^{\text{test}(a)}$

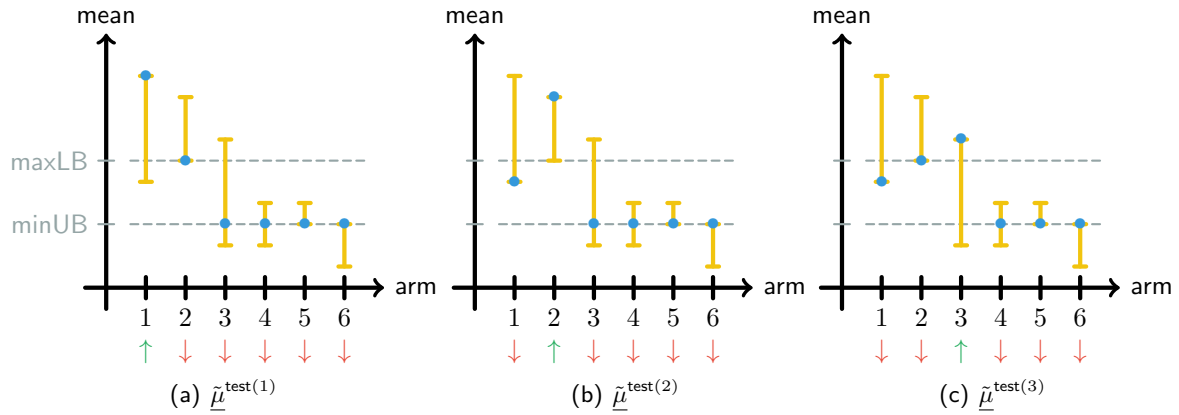


Figure 4.1: List of bandits $(\underline{\tilde{\mu}}^{\text{test}(a)})_{a\in\text{PotentialBest}}$ tried by Algorithm 13 for the example confidence region in yellow. As only the three first arms are above maxLB, they are the potential candidates for being the best arm: PotentialBest $= \{1, 2, 3\}$. For each potential best arm, we associate a unique bandit instance that "maximizes exploration" by putting the mean of that arm as high as possible (↑) and all other arms as lower as possible (↓) while staying above minUB.

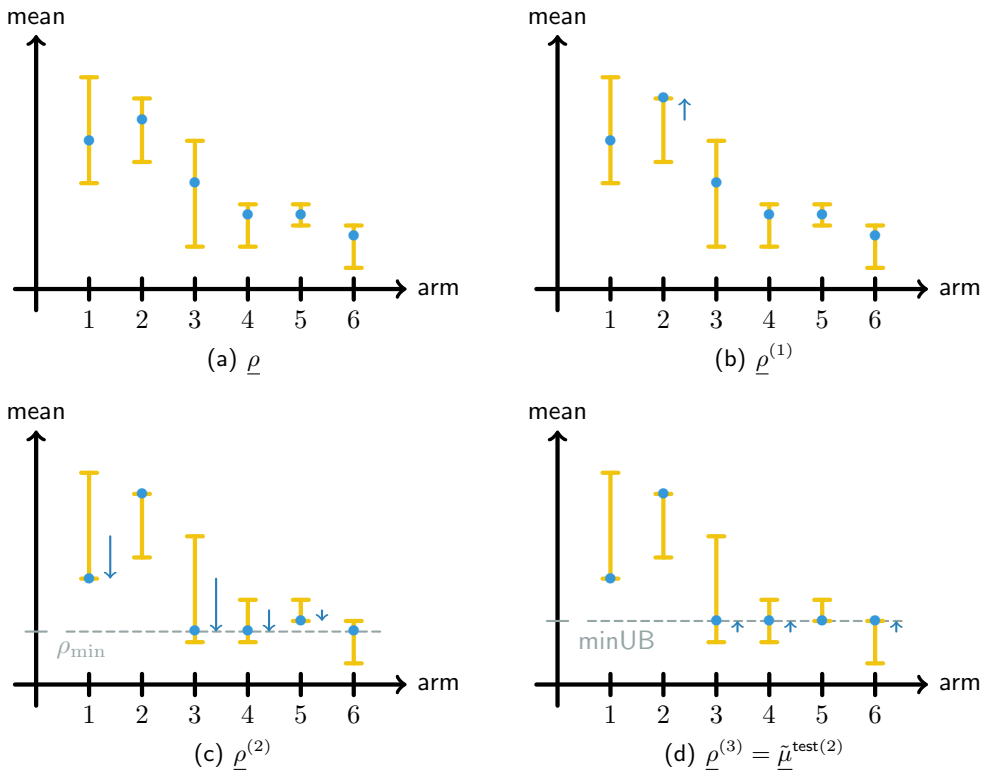


Figure 4.2: Transformations in the proof of Proposition 4.1, for some instance bandit $\underline{\rho}$.

3. Transform $\underline{\rho}^{(2)}$ into $\underline{\rho}^{(3)}$ by increasing all the worst arms to minUB. By Lemma 3.8, one has

$$w_{\min}(\underline{\rho}^{(3)}) \geq w_{\min}(\underline{\rho}^{(2)})\,.$$

Observing that we now get $\underline{\rho}^{(3)} = \underline{\tilde{\mu}}^{\mathsf{test}(a)}$, we proved that

$$w_{\min}(\underline{\rho}) \leq w_{\min}\left(\underline{\tilde{\mu}}^{\mathsf{test}(a)}\right).$$

In conclusion,

$$\max_{\underline{\rho}\in\mathcal{CR}} w_{\min}(\underline{\rho}) = \max_{a\in\mathsf{PotentialBest}} w_{\min}\left(\underline{\tilde{\mu}}^{\mathsf{test}(a)}\right) = w_{\min}(\underline{\tilde{\mu}})\,,$$

where the last inequality comes from the procedure defining $\underline{\tilde{\mu}}$. □

4.2.2. The Strategy

We are now able to introduce our strategy called `Exploration-Biased-Sampling`. Given a risk $\delta \in (0,1)$ and a threshold function $\beta(t,\delta)$, we compute at each time confidence intervals for each μ_a that will ensure $\underline{\mu}$ to belong to each associated confidence region with probability at least $1-\gamma$, where $\gamma \in (0,1)$ is a fixed parameter. We can then ensure enough exploration by biasing the optimal weight vector $\underline{w}(\underline{\mu})$ using Algorithm 13.

Confidence regions. Confidence regions are designed to satisfy two requirements. First, we need products of confidence intervals in order to use Algorithm 13, and then, we will require a time-uniform

confidence guarantee as a key ingredient for the non-asymptotic analysis of `Exploration-Biased-Sampling`. For $\gamma \in (0,1)$ and $t \geq K$, we define

$$\mathcal{CR}_{\underline{\mu}}(t) \overset{\text{def}}{=} \prod_{a \in [K]} \left[\hat{\mu}_a(t) \pm C_{\frac{\gamma}{K}}(N_a(t)) \right], \qquad \text{where} \qquad C_\gamma(s) \overset{\text{def}}{=} 2\sqrt{\frac{\log(\frac{4s}{\gamma})}{s}}. \tag{4.6}$$

The following lemma states a time-uniform γ–confidence guarantee for $\underline{\mu}$.

Lemma 4.2. *Fix $\gamma \in (0,1)$. For all bandit problems $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ (with potentially several optimal arms), it holds that*

$$\mathbb{P}_{\underline{\mu}}\Big(\exists t \geq K : \underline{\mu} \notin \mathcal{CR}_{\underline{\mu}}(t)\Big) \leq \gamma\,.$$

Proof. By union bound we only have to show that, for all $a \in [K]$,

$$\mathbb{P}_{\underline{\mu}}\Big(\exists t \geq K : |\hat{\mu}_a(t) - \mu_a| \geq C_{\frac{\gamma}{K}}(N_a(t))\Big) \leq \frac{\gamma}{K}\,.$$

Let $a \in [K]$. We recall that $N_a(t) \geq 1$ for $t \geq K$, as all arms are observed once at the beginning. Thus, using Equation (4.4), we have

$$\begin{aligned}\mathbb{P}_{\underline{\mu}}\Big(\exists t \geq K : |\hat{\mu}_a(t) - \mu_a| \geq C_{\frac{\gamma}{K}}(N_a(t))\Big) &= \mathbb{P}_{\underline{\mu}}\Big(\exists t \geq K : |\hat{\mu}_{a,N_a(t)} - \mu_a| \geq C_{\frac{\gamma}{K}}(N_a(t))\Big) \\ &= \mathbb{P}_{\underline{\mu}}\Big(\exists n \in \mathbb{N}^* : |\hat{\mu}_{a,n} - \mu_a| \geq C_{\frac{\gamma}{K}}(n)\Big)\,.\end{aligned}$$

By a peeling argument (see, e.g., Boucheron et al., 2013), this leads to

$$\begin{aligned}\mathbb{P}_{\underline{\mu}}\Big(\exists n \in \mathbb{N}^* : |\hat{\mu}_{a,n} - \mu_a| \geq C_{\frac{\gamma}{K}}(n)\Big) &\leq \sum_{k \geq 0} \mathbb{P}_{\underline{\mu}}\left(\exists n \in [2^k, 2^{k+1}] : \left|\frac{1}{n}\sum_{s\in[n]}(X_{a,s} - \mu_a)\right| \geq C_{\frac{\gamma}{K}}(n)\right) \\ &= \sum_{k \geq 0} \mathbb{P}_{\underline{\mu}}\left(\exists n \in [2^k, 2^{k+1}] : \left|\sum_{s\in[n]}X_{a,s} - \mu_a\right| \geq nC_{\frac{\gamma}{K}}(n)\right) \\ &\leq \sum_{k \geq 0} \mathbb{P}_{\underline{\mu}}\left(\exists n \in [0, 2^{k+1}] : \left|\sum_{s\in[n]}X_{a,s} - \mu_a\right| \geq 2^k C_{\frac{\gamma}{K}}(2^k)\right) \\ &\leq 2\sum_{k \geq 0} \exp\left(-\frac{\left(2^k C_{\frac{\gamma}{K}}(2^k)\right)^2}{2 \times 2^{k+1}}\right) \\ &= 2\sum_{k \geq 0} \exp\left(-\log\frac{2^{k+2}K}{\gamma}\right) \\ &= 2\frac{\gamma}{K}\sum_{k \geq 0}\frac{1}{2^{k+2}} \\ &= \frac{\gamma}{K}\,,\end{aligned}$$

where the second inequality is obtained using the fact that $n \mapsto nC_{\frac{\gamma}{K}}(n)$ is non-decreasing and the second inequality is a well-known concentration bound for the sum of sub-Gaussian variables (see Lattimore and Szepesvári, 2020, Theorem 9.2). □

Sampling rule. The sampling rule of `Exploration-Biased-Sampling` is summarized in Algorithm 14. As explained in Garivier and Kaufmann (2016), one can either follow the exploration-biased weights directly (D-tracking) or their cumulative sums (C-tracking). The theoretical results of the strategy will be derived with the use of C-tracking, while we will run the experiments with both options, as D-tracking appears to perform slightly better.

Algorithm 14: `Exploration-Biased-Sampling` sampling rule at step $t > K$

Input: history of observations I_{t-1}
confidence parameter γ
Output: next arm to observe A_t

$\mathcal{CR}_{\underline{\mu}}(t-1) \overset{\text{def}}{=} \prod_{a\in[K]} \left[\hat{\mu}_a(t-1) \pm C_{\frac{\gamma}{K}}\left(N_a(t-1)\right)\right]$

$(\underline{\tilde{\mu}}(t-1), \underline{\tilde{w}}(t-1)) \leftarrow$ `Exploration-Biased-Weights`$(\mathcal{CR}_{\underline{\mu}}(t))$

/* C-tracking */

Choose $A_t \in \operatorname{argmin}_{a\in[K]} N_a(t-1) - \sum_{s\in[t-1]} \tilde{w}_a(s)$

/* D-tracking */

Choose $A_t \in \operatorname{argmin}_{a\in[K]} N_a(t-1) - (t-1)\tilde{w}_a(t-1)$

Stopping and decision rule. Following Garivier and Kaufmann (2016), our stopping rule relies on the statistic

$$Z(t) \overset{\text{def}}{=} \max_{a\in[K]} \min_{b\neq a} Z_{a,b}(t)\,,$$

where $Z_{a,b}(t)$ is the Generalized Likelihood Ratio statistic (see Chernoff, 1959), which is defined, for the standard Gaussian model, as

$$Z_{a,b}(t) \overset{\text{def}}{=} \frac{1}{2}\frac{N_a(t)N_b(t)}{N_a(t)+N_b(t)}\left(\hat{\mu}_a(t)-\hat{\mu}_b(t)\right)^2 \operatorname{sgn}\left(\hat{\mu}_a(t)-\hat{\mu}_b(t)\right). \tag{4.7}$$

The stopping rule consists of stopping the procedure if $Z(t)$ exceeds some threshold $\beta(t,\delta)$, as described in Algorithm 15. See page 50 for more details about the `Global-Likelihood-Ratio` stopping rule. The best empirical arm is then recommended by the strategy: we choose

$$\hat{a}_{\tau_\delta} \in \operatorname*{argmax}_{a\in[K]} \hat{\mu}_a(\tau_\delta)\,.$$

Algorithm 15: `Global-Likelihood-Ratio` stopping rule at step $t > K$

Input: history of observations I_t
threshold function $\beta(t,\delta)$

$Z(t) \leftarrow \max_{a\in[K]} \min_{b\neq a} Z_{a,b}(t)$ // $Z_{a,b}(t)$ is defined in (4.7)

if $Z(t) > \beta(t,\delta)$ **then**
  Stop
else
  Continue

4.3. Theoretical Results

We present the theoretical guarantees of the `Exploration-Biased-Sampling` algorithm.

δ–correctness. Garivier and Kaufmann (2016) proved that, whatever the sampling rule, the use of the `Global-Likelihood-Ratio` stopping rule ensures the δ–correct property for some suitable threshold.

Proposition 4.3. **[Garivier and Kaufmann, 2016, Proposition 12]**
Let $\delta \in (0,1)$ and $\alpha > 1$. There exists a constant $R = R(\alpha, K)$ such that, whatever the sampling rule, using the `Global-Likelihood-Ratio` *stopping rule (Algorithm 15) with threshold*

$$\beta(t,\delta) \overset{\text{def}}{=} \log \frac{Rt^\alpha}{\delta}\,, \tag{4.8}$$

and recommending the best empirical arm ensure that the strategy is δ–correct:

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\mathcal{N}_1}^{[0,1]}, \qquad \mathbb{P}_{\underline{\mu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\mu})) \leq \delta\,.$$

The `Exploration-Biased-Sampling` strategy is hence δ–correct if used with threshold (4.8), whatever the choice of parameter $\gamma \in (0,1)$.

Sufficient exploration. Interestingly, it happens that the choice of the confidence regions (4.6) naturally leads to a minimal exploration rate of the arms of order $\sqrt{t}$.

Lemma 4.4. *Let $\gamma \in (0,1)$. The* `Exploration-Biased-Sampling` *strategy satisfies, for all bandit problems $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$,*

$$\forall t \geq 0, \forall a \in [K], \qquad N_a(t) \geq \frac{2}{K}\sqrt{t} - K\,.$$

What is surprising is that this is exactly the arbitrary rate used by `Track-and-Stop` for forced exploration. The proof of this lemma can be found in Section 4.6.1. Other practical advantages of `Exploration-Biased-Sampling` over `Track-and-Stop` are discussed in Section 4.4.

Non-asymptotic bound. Our main result is to obtain high probability bounds for the sample complexity of `Exploration-Biased-Sampling` in finite horizon.

Theorem 4.5. *Fix $\gamma \in (0,1)$, $\alpha \in [1,2]$, $\eta \in (0,1]$, and let $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$. There exists an event $\mathcal{E}$ of probability at least $1-\gamma$ and $\delta_0 \overset{\text{def}}{=} \delta_0(\underline{\mu}, K, \gamma, \eta, \alpha) > 0$ such that, for all $0 < \delta \leq \delta_0$, algorithm* `Exploration-Biased-Sampling` *with threshold (4.8) satisfies*

$$\forall t > (1+\eta)T(\underline{\mu}) \log\frac{1}{\delta}, \qquad \mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) \leq 2Kt \exp\left(-\frac{t w_{\min}(\underline{\mu})}{4T(\underline{\mu})^2} \frac{1}{\log^{\frac{2}{3}} \frac{1}{\delta}}\right), \tag{4.9}$$

and

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta \mathbb{I}\{\mathcal{E}\}] \leq (1+\eta)T(\underline{\mu}) \log\frac{1}{\delta} + \frac{2^7 K T(\underline{\mu})^4}{w_{\min}(\underline{\mu})^2} \exp\left(-\frac{w_{\min}(\underline{\mu})}{4T(\underline{\mu})} \log^{\frac{1}{3}} \frac{1}{\delta}\right) \log^2 \frac{1}{\delta}\,. \tag{4.10}$$

Section 4.5 will be devoted to the proof of this theorem. The proof will highlight why, contrary to `Track-and-Stop`, the exploration strategy of `Exploration-Biased-Sampling` is adequate for obtaining non-asymptotic bounds. Note that:

- the proof of Theorem 4.5 provides an explicit expression for δ_0,
- the second term of bound (4.10) vanishes when δ decreases to 0, and hence negligible with respect to the first term: the sample complexity is therefore arbitrarily close to lower bound (4.1),
- The dependency on $w_{\min}(\underline{\mu})$ of the bounds might be replaced by the minimal gap $\Delta_{\min}$, as Lemma 4.12 in Section 4.6.1 ensures that

$$\forall \underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}, \qquad w_{\min}(\underline{\mu}) \geq \frac{\Delta_{\min}(\underline{\mu})}{2K}\,.$$

Asymptotic optimality. We additionally prove that, from an asymptotic point of view, our strategy presents the same guarantees as `Track-and-Stop` (see also Theorem 4.17 in Section 4.6.3).

Theorem 4.6. *Let $\gamma \in (0,1)$, $\alpha \in (1, e/2]$. Algorithm* `Exploration-Biased-Sampling` *with threshold* (4.8) *satisfies, for all bandit problems $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log 1/\delta} \leq \alpha T(\underline{\mu})\,. \tag{4.11}$$

The proof of this result can be found in Section 4.6.3.

Remark.

- The multiplicative factor α in bound (4.11) can be avoided using recent proof techniques, as we explain in Section 6.4 (see also Remark 2.11).
- It is worth mentioning that the guarantees of `Exploration-Biased-Sampling` presented in this section hold true not only for Gaussian arms but more generally for 1–sub-Gaussian arms with means in $[0,1]$ (in which case, of course, a better lower bound might hold). Indeed, these proofs only rely on sub-Gaussian deviation bounds.

4.4. Numerical Experiments

In this section, we discuss the behavior and performance of `Exploration-Biased-Sampling` for practical values of risk δ. We propose a comparison with `Track-and-Stop` , `Chernoff-Racing` and `LUCB++`, and begin by recalling a quick description of those strategies:

- the `Track-and-Stop` strategy (see 2.2.5) tracks the optimal weight vector $\underline{w}(\underline{\mu})$ by using the plugin estimate $\underline{w}(\underline{\hat{\mu}}(t))$. Some exploration rate is forced to ensure that bad initial observations do not lead to an under-sampling of some arms. The stopping rule is the same as the one presented for `Exploration-Biased-Sampling`,
- the `Chernoff-Racing` algorithm is an elimination algorithm (see Algorithm 3 in Section 2.2.1): the strategy maintains a list of candidate arms, starting with all arms, and divides the exploration into rounds, during which each arm of the list is observed once. At the end of each round, a decision is made to keep or eliminate the current worst arm from the active set. Several decision rules are possible, we will use the Chernoff rule presented in (Garivier and Kaufmann, 2016), which eliminates the worst empirical arm a_r at the end of round r if

$$Z_{\ell_r, a_r}(t_r) = \frac{r}{4}\big(\hat{\mu}_{\ell_r}(t_r) - \hat{\mu}_{a_r}(t_r)\big)^2 > \beta(t_r, \delta)\,,$$

where ℓ_r (respectively t_r) is the best empirical arm (respectively the number of times steps) at the end of round r.

- the LUCB++ strategy, introduced by Simchowitz et al. (2017) (see also Kalyanakrishnan et al., 2012; Howard et al., 2021) samples two arms at each round: the one with the current best estimate and the one in the remaining arms with the highest optimistic index $U_a(t)$ which is an upper confidence bound[4]:

$$U_a(t) \stackrel{\text{def}}{=} \hat{\mu}_a(t) + \sqrt{\frac{3}{N_a(t)} \log\left(\frac{2K \log(N_a(t))}{\delta}\right)}.$$

  For the fairness of the comparison, we will take the same stopping condition as the strategies `Track-and-Stop` and `Exploration-Biased-Sampling`.

For all strategies, we ran our experiments with the same threshold, given by

$$\beta(t, \delta) = \log\left(\frac{1 + \log(t)}{\delta}\right),$$

and we use, for the `Exploration-Biased-Sampling` strategy, the confidence lengths

$$C_\gamma(s) = \sqrt{\frac{\log(\frac{s}{\gamma})}{s}}.$$

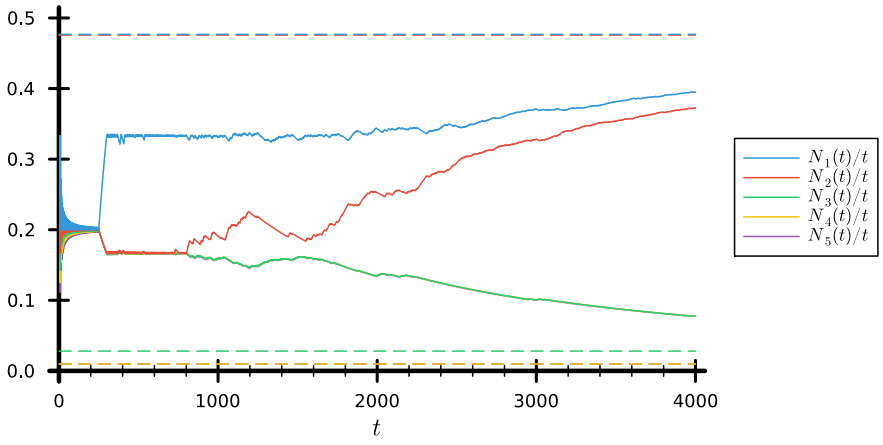
These choices are more aggressive than what the theoretical analysis suggests: yet, empirically, they appear to guarantee the desired failure rate. Using the larger intervals of Section 4.2 would have increased the number of rounds with uniform exploration, and using larger thresholds unnecessarily delays the stopping for all strategies.

Improving the stability of `Track-and-Stop`. In Section 4.1, we highlighted the weaknesses of `Track-and-Stop` (see page 95), especially the forced exploration parameter and the non-interpretable and unstable sampling strategy during the first rounds. In Figures 4.3 and 4.4 we see the improvements of `Exploration-Biased-Sampling` concerning those behaviors. During the first rounds, as for a racing algorithm, a uniform sampling phase is observed as the learner has not collected enough information (the confidence intervals on all arms are not separated), which is the expected behavior. Then the best arms are sampled more and more often, but still in a more cautious way than `Track-and-Stop`. We observe in Figure 4.4 the stability of the sampling strategies compared to `Track-and-Stop` during the first rounds: the targeted weights of `Exploration-Biased-Sampling` are stable and separate from each other cautiously (note that the three last arms still have the same weight at time 1200) whereas, for `Track-and-Stop`, we observe an important variation of the targeted weights with time. As a matter of fact, there is a clear discontinuity each time the estimated best arm changes, as we can see with the red and green arms. We also remark that `Track-and-Stop` uses forced exploration at regular rounds (for the yellow and purple dots), which is unnecessary for `Exploration-Biased-Sampling` as a natural exploration is always performed (Lemma 4.4).

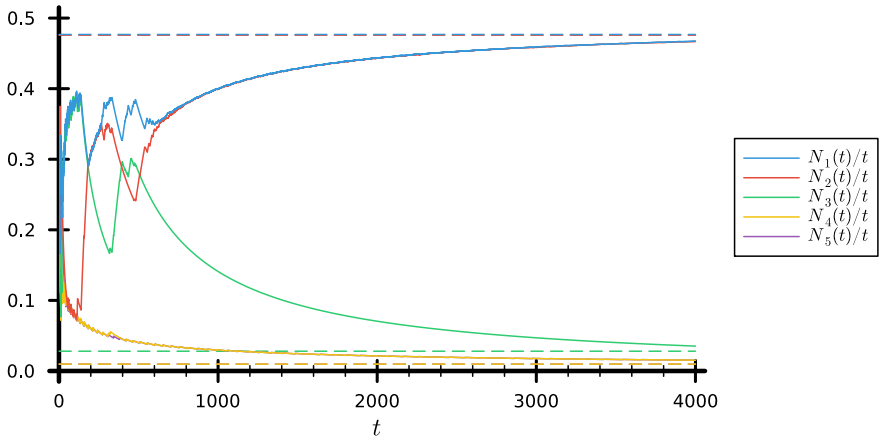
Comparisons of the strategies. The cost of the cautiousness of the algorithm (the exploration-biased weights) is that in terms of pure performance, it takes a little longer for the proportions of draws of `Exploration-Biased-Sampling` to converge to the optimal weight vector. This results in a slightly larger stopping time than `Track-and-Stop` that occurs for all bandit parameters[5]. In other words, `Exploration-Biased-Sampling` does not improve the numerical efficiency of `Track-and-Stop`. This can be observed in Table 4.1, where we present the performances of the strategies with two scenarios and a set of parameters. `Exploration-Biased-Sampling` globally performs

[4] Constant $\sqrt{3}$ appeared to be empirically optimal.

[5] Note that the cautiousness of our strategy is required to obtain the non-asymptotic bounds of Theorem 4.5.



(a) `Exploration-Biased-Sampling`

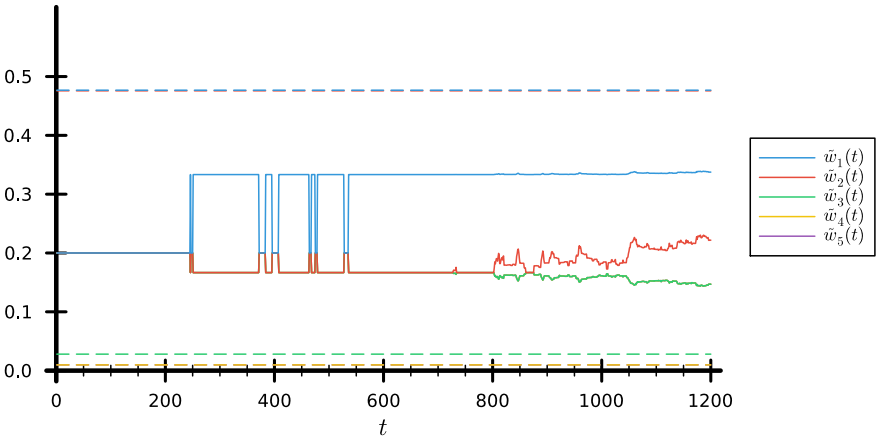


(b) `Track-and-Stop`

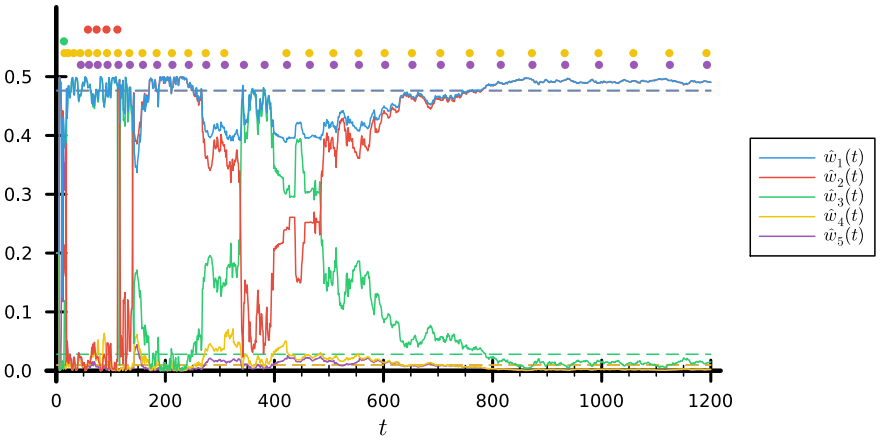
Figure 4.3: Evolution of the sampling frequencies $\frac{N(t)}{t}$ on a single simulation of the strategies `Exploration-Biased-Sampling` and `Track-and-Stop` with D-tracking. The parameters are $\delta = 0.01$, $\gamma = 0.2$ and $\underline{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$. The values of $\underline{w}(\underline{\mu}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ are dashed.

Table 4.1: Empirical expected number of draws $\mathbb{E}_{\underline{\mu}}[\tau_\delta]$, averaged over 1000 experiments, with $\underline{\mu}^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$, $\underline{w}(\underline{\mu}^{(1)}) = (0.477, 0.476, 0.028, 0.010, 0.010)$ and $\underline{\mu}^{(2)} = (0.9, 0.5, 0.45, 0.4)$, $\underline{w}(\underline{\mu}^{(2)}) = (0.375, 0.286, 0.195, 0.144)$.

Instance	δ	γ	$T\,\mathrm{kl}(\delta, 1-\delta)$	EBS-C	TaS-C	EBS-D	TaS-D	CR	LUCB++
$\underline{\mu}^{(1)}$	0.1	0.05	1476	4727	3597	4191	3477	**3124**	3353
$\underline{\mu}^{(1)}$	0.01	0.05	3782	7363	5664	6330	5584	**5419**	5549
$\underline{\mu}^{(1)}$	0.01	0.2	3782	7090	5664	6136	5584	5419	**5372**
$\underline{\mu}^{(1)}$	10^{-5}	0.2	9669	13801	12181	12376	**11439**	11557	11644
$\underline{\mu}^{(2)}$	0.1	0.05	135	476	367	470	**322**	405	365
$\underline{\mu}^{(2)}$	0.01	0.05	347	708	588	699	**485**	542	565



(a) `Exploration-Biased-Sampling`



(b) `Track-and-Stop`

Figure 4.4: Evolution of the Targeted Weights $\underline{\tilde{w}}(t)$ (respectively $\underline{\hat{w}}(t)$) During the First 1200 Rounds on a Simulation of `Exploration-Biased-Sampling` (respectively `Track-and-Stop`). ($\delta = 0.01$, $\gamma = 0.2$, $\underline{\mu} = (0.9, 0.8, 0.6, 0.4, 0.4)$)

correctly but we see that the other strategies are always a little more efficient. Note that when increasing γ, the confidence intervals reduce so that the targeted weights are closer to $\underline{w}$, improving the performance of the algorithm. For similar reasons, the initial cautiousness of the strategy disappears in the long term, thus when δ is very small the relative performance of `Track-and-Stop` and `Exploration-Biased-Sampling` gets closer. Of course, `Exploration-Biased-Sampling` overperforms `Chernoff-Racing` in the long run when the optimal weight vector is far from the sampling proportions of `Chernoff-Racing` (e.g., when $w_1 \gg w_2$).

`Chernoff-Racing` shows great performance with both $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$. This strategy samples the two last arms of the race equally often, thus can be optimal only when $\underline{w}(\underline{\mu})$ has its two highest components of similar value, e.g. when the two best arms are well separated from the others: this is the case of bandit $\underline{\mu}^{(1)}$. For $\underline{\mu}^{(2)}$, any strategy performs well as the problem is easy. However, `Chernoff-Racing` (whose theoretical analysis remains to be written) leads to a few more misidentifications in our experiments that might be linked to the stopping rule we chose here; for fairness reasons, it was taken as identical to that of the other algorithms. `LUCB++` presents similar performance with `Chernoff-Racing`, which can be explained by the similar behavior of the strategies: `LUCB++` samples half of the time the best arm asymptotically, and the worst arms are eliminated one by one once their indexes fall under the two best estimates.

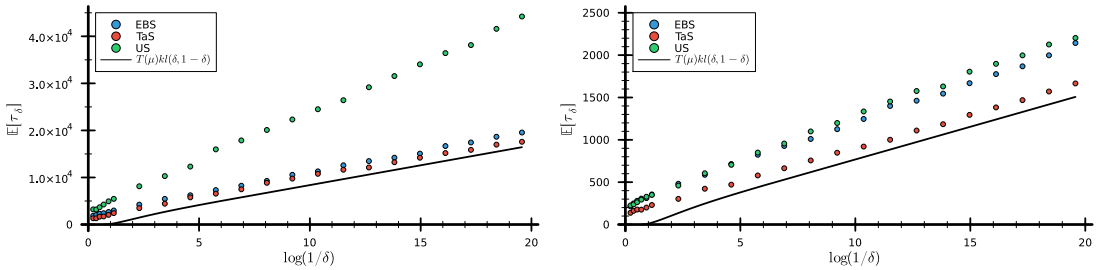
Finally, note that D-tracking shows better performance than C-tracking, either for `Exploration-Biased-Sampling` and `Track-and-Stop`: D-tracking indeed benefits directly from the current estimate of $\underline{\mu}$ (thus the empirical proportions of draws converge faster to $\underline{w}(\underline{\mu})$), while the impact is diluted in time with C-tracking. However, we did not prove theoretical guarantees for D-tracking.

Dependency on the confidence parameter δ. We present numerical experiments to compare the dependency on parameter δ of `Exploration-Biased-Sampling`, `Track-and-Stop` and `Uniform-Sampling` (that samples arms uniformly).

In Figure 4.5, we plot for each strategy and several bandit parameters the estimate of $\mathbb{E}_{\underline{\mu}}[\tau_\delta]$ for different values of δ (using the same threshold β as in the experiments of Section 4 and $\gamma = 0.1$ for `Exploration-Biased-Sampling`). We also plot in black the lower bound of Garivier and Kaufmann (2016), which is of order $T(\underline{\mu}) \log \frac{1}{\delta}$ when δ goes to 0.

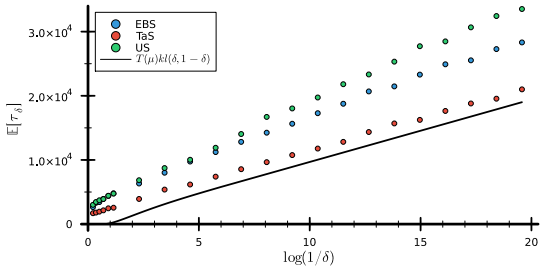
In terms of performance, we observe that `Exploration-Biased-Sampling` is always between `Uniform-Sampling` and `Track-and-Stop`. More precisely, there are different behaviors:

- For some problems (like for instance $\underline{\mu}^{(1)}$), `Exploration-Biased-Sampling` behaves almost like (but always a little worse than) `Track-and-Stop`. For this instance, the uniform sampling phase of `Exploration-Biased-Sampling` is relatively small compared to the required number of samples so that `Exploration-Biased-Sampling` has time to shrink its confidence regions close to parameter $\underline{\mu}$ and thus behaves like `Track-and-Stop`.
- When the problem is easier (with large gaps, see $\underline{\mu}^{(2)}$), the sample complexity is very low and the confidence regions of `Exploration-Biased-Sampling` do not have enough time to shrink. It results in a performance close to `Uniform-Sampling`.
- When the problem is difficult (with small gaps, see $\underline{\mu}^{(3)}$), it takes a large number of samples for `Exploration-Biased-Sampling` before leaving the uniform exploration phase, and this results in a behavior close to `Uniform-Sampling` for moderate values of δ. When δ decreases, there is a separation between `Exploration-Biased-Sampling` and `Uniform-Sampling` as more and more simulations reach the non-uniform sampling phase of our strategy. For even smaller values of δ, one can expect that `Exploration-Biased-Sampling` will come closer to `Track-and-Stop` than `Uniform-Sampling`, for the same reasons as before: the confidence regions of `Exploration-Biased-Sampling` have more time to shrink. This is what we observe with bandit instance $\underline{\mu}^{(4)}$, for which `Exploration-Biased-Sampling` behaves like `Uniform-Sampling` for moderate values of δ and like `Track-and-Stop` for small values of δ.

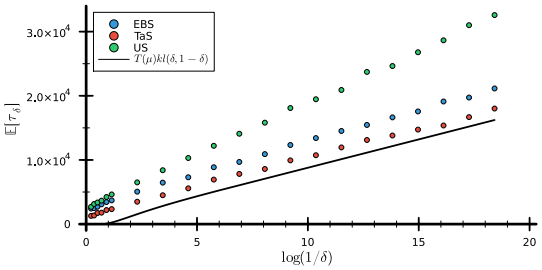


(a) $\underline{\mu}^{(1)} = (0.9, 0.8, 0.6, 0.4, 0.4)$

(b) $\underline{\mu}^{(2)} = (0.9, 0.5, 0.45, 0.4)$



(c) $\underline{\mu}^{(3)} = (0.9, 0.8, 0.75, 0.7)$



(d) $\underline{\mu}^{(4)} = (0.9, 0.8, 0.7, 0.6)$

Figure 4.5: Empirical expected number of draws $\mathbb{E}_{\underline{\mu}}[\tau_\delta]$ as a function of δ, averaged over 500 experiments.

4.5. Proof of the Non-Asymptotic Bounds of Theorem 4.5

The aim of this section is to prove Theorem 4.5. Let $\gamma \in (0,1)$ and $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$. We assume, without loss of generality, that $a^\star(\underline{\mu}) = 1$. We also write, for simplicity, $\underline{\Delta} = \underline{\Delta}(\underline{\mu})$, $\underline{w} = \underline{w}(\underline{\mu})$ and $T = T(\underline{\mu})$.

We recall that the confidence regions are defined in (4.6) by

$$\forall t \geq K, \qquad \mathcal{CR}_{\underline{\mu}}(t) = \prod_{a\in[K]} \Big[\hat{\mu}_a(t) \pm \ell_a(t)\Big]\,,$$

$$\text{where} \qquad \forall a \in [K], \qquad \ell_a(t) \overset{\text{def}}{=} C_{\frac{\gamma}{K}}(N_a(t)) = 2\sqrt{\frac{\log\Big(\frac{4KN_a(t)}{\gamma}\Big)}{N_a(t)}}\,.$$

Let $\mathcal{E}$ denotes the event that $\underline{\mu}$ belongs to all confidence regions:

$$\mathcal{E} = \bigcap_{t=K}^{\tau_\delta} \Big\{\underline{\mu} \in \mathcal{CR}_{\underline{\mu}}(t)\Big\}\,.$$

By the choice of the confidence regions and Lemma 4.2, we have

$$\mathbb{P}_{\underline{\mu}}(\mathcal{E}) \geq 1 - \gamma\,.$$

Furthermore, the design of `Exploration-Biased-Sampling` ensures that, under event $\mathcal{E}$, all arms are observed with some minimal linear rate, specified by Lemma 4.7 and proved in Section 4.6.2.

Lemma 4.7. *On event $\mathcal{E}$, one has:*

$$\forall t \in \mathbb{N}^*, \qquad \min_{a\in[K]} N_a(t) \geq t w_{\min} - K\,. \tag{4.12}$$

Inequality (4.12) implies the more handy bound:

$$\forall t \geq \frac{2K}{w_{\min}}, \qquad \min_{a\in[K]} N_a(t) \geq \frac{t w_{\min}}{2}\,. \tag{4.13}$$

Proof Outline. The proof is organized in 3 steps:

1. We first show that, on event $\mathcal{E}$, the optimal vector $\underline{w}$ and the sampling frequency vector $\underline{N}(t)/t$ are very close for any $t \geq T_1$, where T_1 is a (problem-dependent) constant. To do so, we will make use of the regularity results of Section 3.5 and the fact that the confidence regions shrink with time.
2. Then, we control the event $(\tau_\delta > t) \cap \mathcal{E}$ for $t > T \log\frac{1}{\delta}$ by another event for which we can easily bound the probability using Hoeffding's inequality. This inclusion relies once again on the regularity results of Section 3.5 and on conditions on δ, in particular, we will require to have $T \log\frac{1}{\delta} \geq T_1$ with T_1 obtained at step 1.
3. Finally, we derive the two bounds of the theorem from Hoeffding's inequality and elementary calculations.

The proof uses some technical lemmas introduced and shown in Section 4.6.2.

4.5.1. Step 1: Controlling the Difference between Vectors $\underline{w}$ and $\underline{N}(t)/t$

In this step we assume that event $\mathcal{E}$ occurs.
Let $t \geq \frac{2K}{w_{\min}}$. Equation (4.13) implies that

$$\forall a \in [K], \qquad \ell_a(t) = 2\sqrt{\frac{\log\left(\frac{4N_a(t)K}{\gamma}\right)}{N_a(t)}} \leq L(t) \stackrel{\text{def}}{=} \sqrt{8\frac{\log\left(\frac{4tK}{\gamma}\right)}{tw_{\min}}}\,.$$

$L(t)$ is an arm-independent bound on the half-length of the confidence interval of each μ_a. In other words, $\left\|\tilde{\underline{\mu}}(t) - \underline{\mu}\right\|_\infty \leq L(t)$ as we are on event $\mathcal{E}$. Note that $L(t)$ is deterministic and goes to 0 as t goes to $+\infty$. This control of $\left\|\tilde{\underline{\mu}}(t) - \underline{\mu}\right\|_\infty \leq L(t)$ together with Theorem 3.9 allows to control the difference between $\underline{w}$ and $\tilde{\underline{w}}(t)$ for t large enough, as the following Lemma claims.

Lemma 4.8. *Let*

$$T_0 = \max\left(\frac{224^2}{\Delta_{\min}^2 w_{\min}}\log\left(\frac{2\times 224^2 eK}{\Delta_{\min}^2 w_{\min}\gamma}\right), \frac{2K}{w_{\min}}\right). \tag{4.14}$$

Then for every $t \geq T_0$, one has, introducing $\varepsilon_t = \frac{80L(t)}{\Delta_{\min}}$,

$$\forall a \in [K], \qquad w_a(1-\varepsilon_t) \leq \tilde{w}_a(t) \leq w_a(1+\varepsilon_t)\,. \tag{4.15}$$

Proof. Let $t \geq \frac{2K}{w_{\min}}$ and assume that t is such that $4L(t) < \Delta_{\min}$. On event $\mathcal{E}$, one has

$$\underline{\mu} \in \mathcal{CR}_{\underline{\mu}}(t) = \prod_{a\in[K]}\left[\mu_a^-(t), \overline{\mu}_a(t)\right],$$

hence for any $a \neq 1$:

$$\mu_1^-(t) - \overline{\mu}_a(t) \geq \mu_1 - 2L(t) - \left(\mu_a + 2L(t)\right) \geq \Delta_a - 4L(t) > 0\,,$$

so that the confidence interval for μ_1 is strictly above all other confidence intervals. Hence $\tilde{\underline{\mu}}(t)$ has a unique optimal arm which is arm 1.

For each arm $a \neq 1$, define $\tilde{\Delta}_a(t) \stackrel{\text{def}}{=} \Delta_a(\tilde{\underline{\mu}}(t)) = \tilde{\mu}_1(t) - \tilde{\mu}_a(t)$. Then

$$\tilde{\Delta}_a(t)^2 \leq \left(\Delta_a + 2L(t)\right)^2 = \Delta_a^2\left(1 + \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2}\right) \leq \Delta_a^2\left(1 + \frac{8L(t)}{\Delta_{\min}}\right)$$

$$\text{and} \quad \tilde{\Delta}_a(t)^2 \geq \left(\Delta_a - 2L(t)\right)^2 = \Delta_a^2\left(1 - \frac{4L(t)}{\Delta_a} + \frac{4L(t)^2}{\Delta_a^2}\right) \geq \Delta_a^2\left(1 - \frac{8L(t)}{\Delta_{\min}}\right).$$

If t is such that $\frac{8L(t)}{\Delta_{\min}} \leq \frac{1}{7}$ (this condition is stronger than $4L(t) < \Delta_{\min}$), we can apply Theorem 3.9 which gives

$$\forall a \in [K], \quad w_a(1-\varepsilon_t) \leq \tilde{w}_a(t) \leq w_a(1+\varepsilon_t)\,.$$

It remains to understand when the condition $\frac{8L(t)}{\Delta_{\min}} \leq \frac{1}{7}$ holds. We have:

$$\frac{8L(t)}{\Delta_{\min}} \leq \frac{1}{7} \quad \Longleftrightarrow \quad \frac{\log\left(\frac{4tK}{\gamma}\right)}{t} \leq \frac{\Delta_{\min}^2 w_{\min}}{(7\times 8)^2 \times 8} = \frac{\Delta_{\min}^2 w_{\min}}{2\times 112^2}$$

and this inequality is satisfied, by Lemma 4.15, for

$$t \geq \frac{224^2}{\Delta_{\min}^2 w_{\min}} \log\left(\frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min} \gamma}\right).$$

Combining with the initial condition $t \geq \frac{2K}{w_{\min}}$ leads to the definition of T_0. □

As the C-tracking procedure ensures that each $\frac{N_a(t)}{t}$ is roughly the Cesàro sum of the $(\tilde{w}_a(s))_{0 \leq s \leq t-1}$ (see Lemma 4.14), and as $\varepsilon_t \underset{t \to +\infty}{\longrightarrow} 0$, we are able to control the difference between $\underline{w}$ and $\frac{N(t)}{t}$ after a deterministic time T_1.

Lemma 4.9. *Fix $\eta \in (0,1)$ and let*

$$T_1 \overset{\text{def}}{=} \frac{\max(640^2, 8K)}{\eta^2 \Delta_{\min}^2 w_{\min}^2} \log\left(\frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma}\right). \tag{4.16}$$

Then for any $t \geq T_1$ one has:

$$\forall a \in [K], \qquad w_a(1-\eta) \leq \frac{N_a(t)}{t} \leq w_a(1+\eta).$$

Proof. Let T_0 be defined by Equation (4.14). Let $t > T_0$ and $a \in [K]$. Equation (4.15) of Lemma 4.8 gives:

$$\left|\sum_{s=0}^{t-1} \tilde{w}_a(s) - t w_a\right| \leq \sum_{s=0}^{T_0-1} |\tilde{w}_a(s) - w_a| + \sum_{s=T_0}^{t-1} |\tilde{w}_a(s) - w_a| \leq T_0 + w_a \sum_{s=T_0}^{t-1} \varepsilon_s .$$

By definition of ε_t one has:

$$\sum_{s=T_0}^{t-1} \varepsilon_s = \frac{80\sqrt{8}}{\Delta_{\min}\sqrt{w_{\min}}} \sum_{s=T_0}^{t-1} \sqrt{\frac{\log\left(\frac{4sK}{\gamma}\right)}{s}} \leq \frac{80\sqrt{8}\sqrt{\log\left(\frac{4tK}{\gamma}\right)}}{\Delta_{\min}\sqrt{w_{\min}}} \sum_{s=T_0}^{t-1} \frac{1}{\sqrt{s}} \leq \frac{80\sqrt{8}\sqrt{t\log\left(\frac{4tK}{\gamma}\right)}}{\Delta_{\min}\sqrt{w_{\min}}},$$

so that we have, using Lemma 4.14:

$$\begin{aligned}
\left|\frac{N_a(t)}{t} - w_a\right| &\leq \frac{1}{t}\left[\left|N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s)\right| + \left|\sum_{s=0}^{t-1} \tilde{w}_a(s) - w_a\right|\right] \\
&\leq \frac{K + T_0}{t} + w_a \frac{80\sqrt{8}\sqrt{\log\left(\frac{4tK}{\gamma}\right)}}{\Delta_{\min}\sqrt{w_{\min} t}} \\
&\leq w_a\left(\frac{K + T_0}{t w_{\min}} + \frac{80\sqrt{8}\sqrt{\log\left(\frac{4tK}{\gamma}\right)}}{\Delta_{\min}\sqrt{w_{\min} t}}\right).
\end{aligned}$$

Thus the conclusion of the lemma holds when:

$$\max\left(\frac{K + T_0}{t w_{\min}}, \frac{80\sqrt{8}\sqrt{\log(4tK/\gamma)}}{\Delta_{\min}\sqrt{w_{\min} t}}\right) \leq \frac{\eta}{2},$$

and this inequality is satisfied, using Lemma 4.15, when:

$$t \geq \max\left(\frac{2}{\eta}\frac{K + T_0}{w_{\min}}, \frac{640^2}{\eta^2 \Delta_{\min}^2 w_{\min}} \log\left(\frac{2 \times 640^2 eK}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma}\right)\right).$$

The definition of T_0 implies

$$K + T_0 \leq \frac{4\max(112^2, K)}{\Delta_{\min}^2 w_{\min}} \log\left(\frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min}\gamma}\right),$$

hence the inequality still holds for

$$t \geq \max\left(\frac{8\max(112^2, K)}{\eta\Delta_{\min}^2 w_{\min}^2} \log\left(\frac{2 \times 224^2 eK}{\Delta_{\min}^2 w_{\min}\gamma}\right), \frac{640^2}{\eta^2\Delta_{\min}^2 w_{\min}} \log\left(\frac{2 \times 640^2 eK}{\eta^2\Delta_{\min}^2 w_{\min}\gamma}\right)\right),$$

and T_1 is greater than this lower bound. □

4.5.2. Step 2: a Useful Inclusion of Events

We want to control the event $\{\tau_\delta > t\} \cap \mathcal{E}$ for $t > T \log\frac{1}{\delta}$. For δ small enough, we have the following inclusion of events.

Lemma 4.10. *Fix $\eta \in (0, 0.15]$ and let δ be such that*

$$T \log\frac{1}{\delta} \geq T_1\,, \tag{C1}$$

where T_1 is defined by Equation (4.16), and

$$\log\tfrac{1}{\delta} > \frac{4}{\eta} \log\left(\frac{8eTR^{1/2}}{\eta}\right). \tag{C2}$$

Then for all $C \in (0, 1]$:

$$\forall t \geq (1+C)\frac{(1+\eta)^2}{(1-\eta)^2} T \log\frac{1}{\delta}, \qquad \{\tau_\delta > t\} \cap \mathcal{E} \quad \subseteq \quad \left\{\left\|\underline{\tilde{\mu}}(t) - \underline{\mu}\right\|_\infty \geq \frac{C}{T}\right\} \cap \mathcal{E}\,.$$

Remark. Latter, we will use this Lemma with $C = \frac{1}{\log^{\frac{1}{3}}\frac{1}{\delta}}$.

Proof. Assume in the following that $T\log\frac{1}{\delta} \geq T_1$ and let $t \geq T\log\frac{1}{\delta}$. By definition of T_1 and Lemma 4.9, one has

$$\max_{a\in[K]} \left|\frac{w_a - \frac{N_a(t)}{t}}{w_a}\right| \leq \eta\,.$$

Then using Proposition 3.10 and Equation (4.3):

$$\begin{aligned}
\{\tau_\delta > t\} \cap \mathcal{E} \quad &\subseteq \quad \left\{Z(t) = t\, g\left(\underline{\hat{\mu}}(t), \frac{\underline{N}(t)}{t}\right) \leq \beta(t,\delta)\right\} \cap \mathcal{E} \\
&\subseteq \quad \left\{t\frac{(1-\eta)^2}{1+\eta}\left(g(\underline{\mu}, \underline{w}) - \frac{\left\|\underline{\tilde{\mu}}(t) - \underline{\mu}\right\|_\infty}{2}\right) \leq \beta(t,\delta)\right\} \cap \mathcal{E} \\
&\subseteq \quad \left\{\frac{\left\|\underline{\tilde{\mu}}(t) - \underline{\mu}\right\|_\infty}{2} \geq \frac{1}{T} - \frac{1+\eta}{(1-\eta)^2}\frac{\beta(t,\delta)}{t}\right\} \cap \mathcal{E}\,.
\end{aligned}$$

Consider now

$$f(t) = \frac{1+\eta}{(1-\eta)^2}\frac{\beta(t,\delta)}{t} = \frac{1+\eta}{(1-\eta)^2}\frac{\log(\frac{Rt^\alpha}{\delta})}{t}\,.$$

As $\alpha \leq 2$, one can check that f is decreasing on $(4, +\infty)$. Let us show that

$$\forall C \in (0,1], \qquad f\Big((1+C)\frac{(1+\eta)^2}{(1-\eta)^2} T \log\tfrac{1}{\delta}\Big) \leq \frac{1}{(1+C)T}\,. \tag{4.17}$$

Fix $C \in (0,1]$. As $\alpha \leq 2$ and as $\eta \leq 0.15$ is such that $\frac{(1+\eta)^2}{(1-\eta)^2} \leq 2$, we have:

$$\begin{aligned} f\Big((1+C)\frac{(1+\eta)^2}{(1-\eta)^2} T \log\tfrac{1}{\delta}\Big) &\leq \frac{1+\eta}{(1-\eta)^2} \frac{\log\Big(\frac{R\left(4T\log\frac{1}{\delta}\right)^2}{\delta}\Big)}{(1+C)\frac{(1+\eta)^2}{(1-\eta)^2} T\log\frac{1}{\delta}} \\ &\leq \frac{1}{(1+C)T}\frac{1}{1+\eta}\Big(1 + 2\frac{\log(4R^{1/2}T\log\frac{1}{\delta})}{\log\frac{1}{\delta}}\Big)\,, \end{aligned}$$

hence inequality (4.17) is satisfied if

$$\log(4R^{1/2}T\log\tfrac{1}{\delta}) \leq \frac{\eta}{2}\log\frac{1}{\delta}\,,$$

which is the case, by Lemma 4.15, when:

$$\log\frac{1}{\delta} > \frac{4}{\eta}\log\Big(\frac{8eTR^{1/2}}{\eta}\Big)\,.$$

Finally when inequality (4.17) holds we have for $t \geq (1+C)\frac{(1+\eta)^2}{(1-\eta)^2} T\log\frac{1}{\delta}$:

$$\{\tau_\delta > t\} \cap \mathcal{E} \quad \subseteq \quad \Big\{\big\|\tilde{\underline{\mu}}(t) - \underline{\mu}\big\|_\infty \geq \frac{2}{T} - \frac{2}{(1+C)T}\Big\} \cap \mathcal{E} \quad \subseteq \quad \Big\{\big\|\tilde{\underline{\mu}}(t) - \underline{\mu}\big\|_\infty \geq \frac{C}{T}\Big\} \cap \mathcal{E}$$

where we use $C \leq 1$ in the last inclusion. □

4.5.3. Step 3: Bounding $\mathbb{P}_{\underline{\mu}}\big(\tau_\delta > t \cap \mathcal{E}\big)$ and $\mathbb{E}_{\underline{\mu}}\big[\tau_\delta \mathbb{I}\{\mathcal{E}\}\big]$.

Fix $\eta \in (0,1]$ and assume in the following that conditions (C1) and (C2) of Lemma 4.10 are satisfied with $\eta' = \frac{\eta}{7} \leq 0.15$. We set $\zeta = \frac{(1+\eta')^2}{(1-\eta')^2}$. Let $C \in (0,1]$, $t > (1+C)\zeta T\log\frac{1}{\delta}$ and define

$$\mathcal{E}_t = \Big\{\big\|\tilde{\underline{\mu}}(t) - \underline{\mu}\big\|_\infty \geq \frac{C}{T}\Big\} \bigcap \mathcal{E}\,.$$

Lemmas 4.10 and 4.16 – a consequence of Hoeffding's inequality – (note that Condition (C1) ensures that $t \geq \frac{2K}{w_{\min}}$) give the bound:

$$\mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) \leq \mathbb{P}_{\underline{\mu}}(\mathcal{E}_t) \leq 2Kt\exp\Big(-\frac{tw_{\min}}{4T^2}C^2\Big)\,. \tag{4.18}$$

By taking $C = \frac{1}{\log^{\frac{1}{3}}\frac{1}{\delta}}$, we obtained so far that

$$\forall t > \Big(1 + \frac{1}{\log^{\frac{1}{3}}\frac{1}{\delta}}\Big)\zeta T\log\frac{1}{\delta}, \qquad \mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) \leq 2Kt\exp\Big(-\frac{tw_{\min}}{4T^2}\frac{1}{\log^{\frac{2}{3}}\frac{1}{\delta}}\Big)\,,$$

giving Bound (4.9) as long as

$$\Big(1 + \frac{1}{\log^{\frac{1}{3}}\frac{1}{\delta}}\Big)\zeta \leq 1 + \eta\,.$$

Note that $\zeta \leq 1 + 6\eta'$ as $\eta' \leq 0.15$ so that when

$$\frac{1}{\log^{\frac{1}{3}} \frac{1}{\delta}} \leq \frac{\eta'}{2} \quad \iff \quad \log \frac{1}{\delta} \geq \frac{8 \times 7^3}{\eta^3}\,, \tag{C3}$$

the condition holds as

$$\left(1 + \frac{1}{\log^{\frac{1}{3}} \frac{1}{\delta}}\right)\zeta \leq \left(1 + \frac{\eta'}{2}\right)(1 + 6\eta') \leq 1 + 6.6\eta' \leq 1 + \eta\,.$$

It remains to focus on the bound of $\mathbb{E}_{\underline{\mu}}[\tau_\delta \mathbb{I}\{\mathcal{E}\}]$. Using Equation (4.18) we have:

$$\begin{aligned}\mathbb{E}_{\underline{\mu}}[\tau_\delta \mathbb{I}\{\mathcal{E}\}] &= \sum_{t=0}^{\left\lfloor (1+C)\zeta T \log \frac{1}{\delta} \right\rfloor} \mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) + \sum_{t > (1+C)\zeta T \log \frac{1}{\delta}} \mathbb{P}_{\underline{\mu}}(\tau_\delta > t \cap \mathcal{E}) \\ &\leq (1+C)\zeta T \log \frac{1}{\delta} + 1 + 2K \sum_{t > (1+C)\zeta T \log \frac{1}{\delta}} t \exp\left(-\frac{t w_{\min}}{4T^2} C^2\right).\end{aligned}$$

By defining

$$S(C) = \sum_{t > C\zeta T \log \frac{1}{\delta}} t \exp\left(-\frac{t w_{\min}}{4T^2} C^2\right),$$

we obtain, with some technical calculations (see Section 4.6.2), the following bound.

Lemma 4.11. *One has*

$$S(C) \leq \frac{32T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} C^2 \log \tfrac{1}{\delta}\right)\left(\frac{\log \frac{1}{\delta}}{C^2} + \frac{1}{C^4}\right).$$

Once again, taking $C = \frac{1}{\log^{\frac{1}{3}} \frac{1}{\delta}}$ leads to

$$S(C) \leq \frac{32T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}} \tfrac{1}{\delta}\right)\left(\log^{\frac{5}{3}} \tfrac{1}{\delta} + \log^{\frac{4}{3}} \tfrac{1}{\delta}\right) \leq \frac{64T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}} \tfrac{1}{\delta}\right) \log^2 \frac{1}{\delta}\,,$$

thus

$$\mathbb{E}_{\underline{\mu}}[\tau_\delta \mathbb{I}\{\mathcal{E}\}] \leq \zeta\left(1 + \frac{1}{\log^{\frac{1}{3}} \frac{1}{\delta}}\right) T \log \frac{1}{\delta} + 1 + \frac{2^7 K T^4}{w_{\min}^2} \exp\left(-\frac{w_{\min}}{4T} \log^{\frac{1}{3}} \tfrac{1}{\delta}\right) \log^2 \frac{1}{\delta}\,.$$

Under Condition (C3) we get

$$\zeta\left(1 + \frac{1}{\log^{\frac{1}{3}} \frac{1}{\delta}}\right) T \log \frac{1}{\delta} + 1 \leq (1 + 6.6\eta') T \log \frac{1}{\delta} + 1 \leq (1+\eta) T \log \frac{1}{\delta}\,,$$

and obtain the Bound (4.10) claimed in the theorem.

Combining conditions (C1), (C2) and (C3) together, one can define δ_0 satisfying:

$$\log \frac{1}{\delta_0} \geq \frac{7^3 \times \max(2 \times 160^2, K)}{\eta^3 \Delta_{\min} w_{\min}^2} \log\left(\frac{7^2 \times 2 \times 640^2 e K R^{1/2}}{\eta^2 \Delta_{\min}^2 w_{\min} \gamma}\right),$$

with some simplifications allowed by Equation (3.20) of Proposition 3.4.

4.6. Technical Results

4.6.1. Proof of Lemma 4.4

We will prove the lemma using two supporting results. The following lemma gives a lower bound of $w_{\min}(\underline{\mu})$ in terms of the minimal gap $\Delta_{\min}(\underline{\mu})$, based on the study of optimization problem (4.2) in Chapter 3.

Lemma 4.12. *For all $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$ with a unique optimal arm, one has*

$$w_{\min}(\underline{\mu}) \geq \frac{\Delta_{\min}(\underline{\mu})}{2K}\,.$$

Proof. Let $\underline{w} = \underline{w}(\underline{\mu})$, $w_{\min} = w_{\min}(\underline{\mu})$ and $\underline{\Delta} = \underline{\Delta}(\underline{\mu})$. We have, by Equation (3.10), recalling that the weight $w_{\min}$ is the optimal weight of the worst arm(s) of $\underline{\mu}$, i.e., the arm(s) with highest gap(s),

$$w_{\min} = \frac{w_{\max}}{r\Delta_{\max} - 1}\,.$$

Applying inequalities (3.18) and (3.19), and using the fact that $\Delta_{\max} \leq 1$ as means belong to $[0,1]$,

$$w_{\min} \geq \frac{1}{\sqrt{K-1}+1} \cdot \frac{1}{\frac{\sqrt{K-1}+1}{\Delta_{\min}}\Delta_{\max} - 1} \geq \frac{\Delta_{\min}}{(\sqrt{K-1}+1)^2} \geq \frac{\Delta_{\min}}{2K}\,.$$

□

The next lemma provides a lower bound on the minimal gap of the optimistic bandit $\underline{\tilde{\mu}}$ computed by Algorithm 13.

Lemma 4.13. *Let $\mathcal{CR} = \prod_{a\in[K]} [\mu_a^-, \mu_a^+]$ be a confidence region with $\mu_a^- < \mu_a^+$ for all $a \in [K]$, and assume that*

$$\max_{a\in[K]} \mu_a^- = \mathit{maxLB} > \mathit{minUB} = \min_{a\in[K]} \mu_a^+\,.$$

Then, if $(\underline{\tilde{\mu}}, \underline{v}) \leftarrow$ `Exploration-Biased-Weights`$(\mathcal{CR})$, we get

$$\Delta_{\min}(\underline{\tilde{\mu}}) \geq \min_{a\in[K]} \mu_a^+ - \mu_a^-\,.$$

Proof. We proceed by contradiction: let us assume that $\underline{\tilde{\mu}}$ is such that

$$\Delta_{\min}(\underline{\tilde{\mu}}) < \min_{a\in[K]} \mu_a^+ - \mu_a^-\,.$$

By the two hypotheses and the algorithm's procedure, it is clear that $\underline{\tilde{\mu}}$ has a unique best arm. Without loss of generality let us arrange the arms so that $\tilde{\mu}_1 > \tilde{\mu}_2 \geq \tilde{\mu}_3 \geq \cdots \geq \tilde{\mu}_K$. Note that $\Delta_{\min}(\underline{\tilde{\mu}}) = \tilde{\mu}_1 - \tilde{\mu}_2$.

As 1 is the best arm, once again the algorithm's procedure ensures that $\tilde{\mu}_1 = \mu_1^+$. In addition, our assumption implies $\Delta_{\min}(\underline{\tilde{\mu}}) < \mu_1^+ - \mu_1^-$, giving $\tilde{\mu}_2 > \mu_1^-$. Recall that $\tilde{\mu}_2 = \max(\mu_2^-, \mathrm{minUB})$, so we split our analysis into the two possible cases:

- if $\tilde{\mu}_2 = \mu_2^-$, then we cannot have $\mu_2^+ \leq \mu_1^+ = \tilde{\mu}_1$ otherwise $\Delta_{\min}(\underline{\tilde{\mu}}) > \mu_2^+ - \mu_2^-$, which is impossible.
  Then $\mu_2^+ > \mu_1^+$. By defining $\underline{\rho} = (\tilde{\mu}_2, \mu_2^+, \tilde{\mu}_3, \dots, \tilde{\mu}_K)$, one has $\underline{\rho} \in \mathcal{CR}$ and $w_{\min}(\underline{\rho}) > w_{\min}(\underline{\tilde{\mu}})$ by Lemma 3.6. Thus $\underline{\tilde{\mu}}$ cannot maximize $w_{\min}$ over $\mathcal{CR}$ which is in contradiction with Proposition 4.1.
- if $\tilde{\mu}_2 = \text{minUB}$, then $\tilde{\mu}_2 = \tilde{\mu}_3 = \cdots = \tilde{\mu}_K$ and thus all confidence intervals share a common point equal to $\tilde{\mu}_2$ (recall that $\tilde{\mu}_2 \in [\mu_1^-, \mu_1^+]$), which is a contradiction with maxLB $>$ minUB.

□

We can now prove Lemma 4.4.

Proof of Lemma 4.4. Let $t \geq 0$. We want to lower bound $\tilde{w}_{\min}(t)$.

- If at time t one has $\underline{\tilde{w}}(t) = (\frac{1}{K}, \dots, \frac{1}{K})$, that is, all confidence intervals share a common value, then $\tilde{w}_{\min}(t) = \frac{1}{K}$.
- Otherwise, by the construction of Algorithms 13 and 14 we know that $t \geq K$ and the confidence region $\mathcal{CR}(t)$ is such that at least two confidence intervals are separated. In that case, the optimistic bandit $\underline{\tilde{\mu}}(t)$ has a unique optimal arm, hence by Lemma 4.12

$$\tilde{w}_{\min}(t) \geq \frac{\tilde{\Delta}_{\min}(t)}{2K}\,.$$

  Applying Lemma 4.13 leads to (recalling that $N_a(t) \geq 1$ as all arms are pulled once at the beginning)

$$\tilde{\Delta}_{\min}^{(t)} \geq \min_{a\in[K]} 2\ell_a(t) \geq 4 \min_{a\in[K]} \sqrt{\frac{\log\left(\frac{4N_a(t)K}{\gamma}\right)}{N_a(t)}} \geq 4\sqrt{\frac{\log(\frac{4K}{\gamma})}{t}} \geq 4\sqrt{\frac{\log 8}{t}} \geq \frac{4}{\sqrt{t}}\,.$$

  Combining the two last equations, we obtain

$$\tilde{w}_{\min}(t) \geq \frac{2}{K}\frac{1}{\sqrt{t}}\,.$$

Hence, in both cases, we get:

$$\forall t \geq 0, \qquad \tilde{w}_{\min}(t) \geq \min\left(\frac{2}{K}\frac{1}{\sqrt{t}}, \frac{1}{K}\right) \geq \frac{1}{K}\frac{1}{\sqrt{t}}\,.$$

Using Lemma 4.14, this implies that for all $t \geq 0$ and $a \in [K]$

$$N_a(t) \geq \sum_{s=0}^{t-1} \tilde{w}_a(s) - (K-1) \geq \sum_{s=2}^{t-1} \tilde{w}_{\min}(s) - K \geq \frac{1}{K}\sum_{s=2}^{t-1}\frac{1}{\sqrt{s}} - K \geq \frac{1}{K}\int_1^t \frac{\mathrm{d}s}{\sqrt{s}} - K \geq \frac{2}{K}\sqrt{t} - K\,.$$

□

4.6.2. Technical Details for the Proof of Theorem 4.5

Proof of Lemma 4.7. We will use the following deterministic Lemma.

Lemma 4.14. *One has:*

$$\forall t > 0, \forall a \in [K], \qquad \left|N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s)\right| \leq K - 1\,.$$

Proof. Apply Garivier and Kaufmann (2016, Lemma 15) with $p(s) = \underline{\tilde{w}}(s)$. □

Proof of Lemma 4.7. The claim is true for $t \leq K$. Otherwise, fix $t > K$ and $a \in [K]$. For all $0 \leq s \leq K-1$, one has $\tilde{w}_a(s) = \frac{1}{K}$ by convention[6], and thus $\tilde{w}_a(s) \geq w_{\min}$. As $\underline{\mu} \in \mathcal{CR}_{\underline{\mu}}(s)$ for all $K \leq s \leq t$ on event $\mathcal{E}$, Proposition 4.1 ensures that

$$\forall K \leq s \leq t, \qquad \tilde{w}_a(s) \geq \tilde{w}_{\min}(s) = \max_{\underline{\rho} \in \mathcal{CR}_{\underline{\mu}}(s)} w_{\min}(\underline{\rho}) \geq w_{\min}\,.$$

Hence, by Lemma 4.14

$$N_a(t) \geq \sum_{s=0}^{t-1} \tilde{w}_a(s) - (K-1) \geq t w_{\min} - (K-1) \geq t w_{\min} - K\,.$$

□

A Technical Lemma. The following result is a direct consequence of Garivier and Kaufmann (2016, Lemma 18).

Lemma 4.15. *For any $c_1, c_2 > 0$,*

$$x = \frac{2}{c_1} \log\Big(\frac{c_2 e}{c_1}\Big)$$

is such that $c_1 x \geq \log(c_2 x)$.

Deviation Bound. We prove the following simple consequence of Hoeffding's inequality.

Lemma 4.16. *For all $t \geq \frac{2K}{w_{\min}}$ and $x > 0$, one has*

$$\mathbb{P}_{\underline{\mu}}\Big(\max_{a \in [K]} |\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E}\Big) \leq 2Kt \exp\Big(-\frac{t w_{\min}}{4} x^2\Big)\,.$$

Proof. Fix $t \geq \frac{2K}{w_{\min}}$ and $x > 0$. With $T = \frac{t w_{\min}}{2}$, we get, using Equations (4.13) and (4.4),

$$\begin{aligned}
\forall a \in [K], \qquad \mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E}\Big) &= \sum_{s=T}^{t} \mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_a(t) - \mu_a| > x \cap \mathcal{E} \cap N_a(t) = s\Big) \\
&\leq \sum_{s=T}^{t} \mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_{a,s} - \mu_a| > x\Big) \\
&\leq \sum_{s=T}^{t} 2\exp\Big(-\frac{s}{2}x^2\Big) \\
&\leq 2t \exp\Big(-\frac{T}{2}x^2\Big)\,,
\end{aligned}$$

where the second inequality uses Hoeffding's inequality (2.4). The result follows by union bound. □

Proof of Lemma 4.11. We have

$$S(C) = \sum_{t > (1+C)\zeta T \log\frac{1}{\delta}} t \exp\Big(-\frac{t w_{\min}}{4T^2} C^2\Big) = \sum_{t > B} f(t)$$

[6] As all arms are drawn once during the K first rounds, the only request is $\sum_{s=0}^{K-1} \tilde{w}_a(s) = 1$.

where $f : t \mapsto t\exp(-At)$, $A = \frac{w_{\min}}{4T^2}C^2$ and $B = (1+C)\zeta T \log\frac{1}{\delta}$. f is increasing until $1/A$ and then decreasing. Let $n_0 = \left\lfloor \frac{1}{A} \right\rfloor$. We will show that $S(C) \leq 2\int_B^{+\infty} f(t)\ \mathrm{d}t$.

- If $B > n_0$ then f is decreasing on $[B, +\infty[$ and one has $S(C) \leq \int_B^{+\infty} f(t)\ \mathrm{d}t$.
- Otherwise, one has:

$$\begin{aligned} S(C) &= \sum_{t=\lceil B \rceil}^{n_0-1} f(t) + f(n_0) + f(n_0+1) + \sum_{t>n_0+1} f(t) \\ &\leq \sum_{t=\lceil B \rceil}^{n_0-1} \int_t^{t+1} f(t)\ \mathrm{d}t + f(n_0) + f(n_0+1) + \sum_{t>n_0+1} \int_{t-1}^{t} f(t)\ \mathrm{d}t \\ &\leq \int_{\lceil B \rceil}^{+\infty} f(t)\ \mathrm{d}t + f(n_0) + f(n_0+1) \end{aligned}$$

where in the second inequality, we use the increasing (respectively the decreasing) of f on $[B, n_0]$ (respectively on $[n_0+1, +\infty]$). The result will be true if

$$f(n_0) + f(n_0+1) \leq \int_B^{+\infty} f(t)\ \mathrm{d}t\,.$$

We have:

$$\begin{aligned} f(n_0) + f(n_0+1) &= \left\lfloor \frac{1}{A} \right\rfloor \mathrm{e}^{-A\left\lfloor \frac{1}{A} \right\rfloor} + \left\lceil \frac{1}{A} \right\rceil \mathrm{e}^{-A\left\lceil \frac{1}{A} \right\rceil} \\ &\leq \Big(\left\lfloor \frac{1}{A} \right\rfloor + \left\lceil \frac{1}{A} \right\rceil \Big) \mathrm{e}^{-A\left\lfloor \frac{1}{A} \right\rfloor} \\ &\leq \Big(\left\lfloor \frac{1}{A} \right\rfloor \frac{1}{A} + \frac{1}{A^2} \Big) \mathrm{e}^{-A\left\lfloor \frac{1}{A} \right\rfloor} && \text{as } A < \frac{1}{2} \\ &= \int_{\left\lfloor \frac{1}{A} \right\rfloor}^{+\infty} f(t)\ \mathrm{d}t \leq \int_B^{+\infty} f(t)\ \mathrm{d}t && \text{as } B \leq \left\lfloor \frac{1}{A} \right\rfloor = n_0\,. \end{aligned}$$

where in the last inequality, we used the simple calculation

$$\int_Y^{+\infty} t\exp(-tX)\mathrm{d}t = \exp(-YX)\Big(\frac{Y}{X} + \frac{1}{X^2}\Big)$$

for $X, Y > 0$.

In both cases, we have:

$$S(C) \leq 2\int_{(1+C)\zeta T \log\frac{1}{\delta}}^{\infty} t\exp\Big(-\frac{tw_{\min}}{4T^2}C^2\Big)\ \mathrm{d}t$$

and using the same calculation as before

$$S(C) \leq 2\exp\Big(-\frac{\zeta w_{\min}}{4T}(1+C)C^2\log\tfrac{1}{\delta}\Big)\Big(\frac{4(1+C)\zeta T^3}{w_{\min}}\frac{\log\frac{1}{\delta}}{C^2} + \frac{16T^4}{w_{\min}^2}\frac{1}{C^4}\Big)\,.$$

Bounding $C \in (0,1]$ and $\zeta \in [1,2]$ (remind that $\zeta \leq 1 + 6\eta'$):

$$\begin{aligned} S(C) &\leq 2\exp\Big(-\frac{w_{\min}}{4T}C^2\log\tfrac{1}{\delta}\Big)\Big(\frac{16T^3}{w_{\min}}\frac{\log\frac{1}{\delta}}{C^2} + \frac{16T^4}{w_{\min}^2}\frac{1}{C^4}\Big) \\ &\leq \frac{32T^4}{w_{\min}^2}\exp\Big(-\frac{w_{\min}}{4T}C^2\log\tfrac{1}{\delta}\Big)\Big(\frac{\log\frac{1}{\delta}}{C^2} + \frac{1}{C^4}\Big)\,. \end{aligned}$$

□

4.6.3. Almost Sure Asymptotic Bound

Almost Sure Asymptotic Bound. `Exploration-Biased-Sampling` satisfies the following almost sure asymptotic bound.

Theorem 4.17. *Let $\gamma \in (0,1)$ and $\alpha \in [1, e/2]$. Algorithm* `Exploration-Biased-Sampling` *with threshold (4.8) satisfies*

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\mathcal{N}_1}^{[0,1]}, \qquad \limsup_{\delta \to 0} \frac{\tau_\delta}{\log \frac{1}{\delta}} \leq \alpha T(\underline{\mu}) \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.}\,.$$

This result was obtained for the `Track-and-Stop` algorithm by Garivier and Kaufmann (2016, Proposition 13). The adaptation to `Exploration-Biased-Sampling` is straightforward, as soon as we prove the almost sure convergence of the empirical means and frequencies of pulls.

Proposition 4.18. *For all $\gamma \in (0,1)$, the sampling rule of* `Exploration-Biased-Sampling` *satisfies, for all $\underline{\mu} \in \mathcal{D}_{\mathcal{N}_1}^{[0,1]}$,*

$$\lim_{t \to +\infty} \underline{\hat{\mu}}(t) = \underline{\mu} \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.} \qquad \text{and} \qquad \lim_{t \to +\infty} \frac{N(t)}{t} = w(\underline{\mu}) \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.}\,.$$

Proof. The first convergence is a simple application of the law of large numbers, given that the number of pulls of all arms diverges thanks to Lemma 4.4. Note also that, as

$$\forall a \in [K], \qquad \left|\tilde{\mu}_a(t) - \hat{\mu}_a(t)\right| \leq C_{\frac{\gamma}{K}}(N_a(t)) = 2\sqrt{\frac{\log\left(\frac{4N_a(t)K}{\gamma}\right)}{N_a(t)}} \underset{t \to +\infty}{\longrightarrow} 0\,,$$

we have that $\underline{\tilde{\mu}}(t)$ and $\underline{\hat{\mu}}(t)$ get the same limit:

$$\lim_{t \to +\infty} \underline{\tilde{\mu}}(t) = \underline{\mu} \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.}.$$

Thus, by continuity of the optimal weights $\underline{w}$ at $\underline{\mu}$:

$$\lim_{t \to +\infty} \underline{\tilde{w}}(t) = \underline{w}(\underline{\mu}) \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.}\,.$$

Finally, for all $t \in \mathbb{N}^*$ and $a \in [K]$, we have:

$$\begin{aligned}\left|\frac{N_a(t)}{t} - w_a(\underline{\mu})\right| &\leq \frac{1}{t}\left|N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s)\right| + \left|\frac{1}{t}\sum_{s=0}^{t-1}(\tilde{w}_a(s) - w_a(\underline{\mu}))\right| \\ &\leq \frac{K-1}{t} + \left|\frac{1}{t}\sum_{s=0}^{t-1}(\tilde{w}_a(s) - w_a(\underline{\mu}))\right| \\ &\underset{t \to +\infty}{\longrightarrow} 0 \quad \mathbb{P}_{\underline{\mu}}\text{-a.s.}\,,\end{aligned}$$

using Lemma 4.14 for the second inequality, and Cesàro Lemma for the convergence. □

Asymptotic optimality. We now prove Theorem 4.6 on the asymptotic optimality of our strategy. Once again this is a direct adaptation of Garivier and Kaufmann (2016, Theorem 14). One can apply the original proof as long as the two lemmas shown in this section are satisfied.

Proof. Let $T \geq 0$. We get, for all $t \geq \sqrt{T} = h(T)^2$ and $a \in [K]$,

$$\begin{aligned}
\left|\frac{N_a(t)}{t} - w_a(\underline{\mu})\right| &\leq \frac{1}{t}\left|N_a(t) - \sum_{s=0}^{t-1} \tilde{w}_a(s)\right| + \left|\frac{1}{t}\sum_{s=0}^{t-1}(\tilde{w}_a(s) - w_a(\underline{\mu}))\right| \\
&\leq \frac{K-1}{t} + \frac{h(T)}{t} + \left|\frac{1}{t}\sum_{s=h(T)}^{t-1}(\tilde{w}_a(s) - w_a(\underline{\mu}))\right| && \text{by Lemma 4.14} \\
&\leq \frac{K-1}{T^{1/2}} + \frac{1}{T^{1/4}} + \varepsilon && \text{by definition of } \mathcal{E}_T \\
&\leq \frac{K}{T^{1/4}} + \varepsilon\,, \leq 3\varepsilon\,,
\end{aligned}$$

where we used Lemma 4.14 in the second inequality, and the definition of $\mathcal{E}_T$ for the third inequality. This gives the result whenever $T \geq T_\varepsilon \stackrel{\text{def}}{=} \left(\frac{K}{2\varepsilon}\right)^4$. □

4.7. Conclusion

We introduced `Exploration-Biased-Sampling`, a new strategy for the problem of best arm identification with fixed confidence. In addition to asymptotic optimal results, we proved non-asymptotic bounds for this strategy in the case of (sub-)Gaussian bandits. Those finite risk bounds were made possible by a new analysis of the sample complexity optimization problem presented in Chapter 3, and by the design of our strategy which tackles some shortcomings of `Track-and-Stop`: the procedure ensures exploration in an unforced way and stabilizes the sampling strategy, observing uniformly before having a high certainty that one arm is better than another.

Improving the guarantees of `Exploration-Biased-Sampling`. Although our new strategy enjoys interesting stability properties, it is always a bit worse than `Track-and-Stop` in terms of pure performance (sample complexity). This is a consequence of the biased tracking of the strategy which implies a slower convergence of the pulling frequencies than `Track-and-Stop`, but was necessary to obtain our non-asymtotic bound. A future direction of research might consist in carefully modifying the exploration mechanism of `Exploration-Biased-Sampling` in order to obtain a more efficient strategy that still benefits from a finite risk analysis. The non-asymptotic guarantees of Theorem 4.5 also come with a few limitations, which leaves room for improvement:

- The finite risk bound is given on an event $\mathcal{E}$ of high probability (depending on the external parameter γ for the confidence regions). What happens when $\mathcal{E}$ does not occur?
- The analysis is valid for values of confidence δ that are in practice extremely small. Can we analyze it for more moderate values of δ?

Non-Gaussian models. It would be interesting but it remains out of reach to generalize this approach to non-Gaussian models: this requires extending our results on the sample-complexity optimization problem, a technically challenging task for which the simple and clean arguments developed here are likely to be replaced by much more involved derivations if this is possible. In addition, it will be necessary to modify the confidence intervals on the arm means in a way that ensures exploration. Another direction of improvement will be to investigate if similar analysis and strategies are possible for the problem of ε-best arm identification.

CHAPTER 5

A Non-Parametric Theory of Fixed-Budget Best-Arm Identification

In this chapter, we study non-parametric generalizations of existing bounds in fixed-budget best-arm identification. We consider general models $\mathcal{D}$ for distributions over the arms; an overarching example is the model $\mathcal{D} = \mathcal{P}[0,1]$ of all probability distributions over $[0,1]$. We propose upper bounds on the average log-probability of misidentifying the optimal arm based on information-theoretic quantities that we name $\mathcal{L}^{<}_{\inf}(\,\cdot\,,\nu)$ and $\mathcal{L}^{>}_{\inf}(\,\cdot\,,\nu)$ and that correspond to infima over Kullback-Leibler divergences between some distributions in $\mathcal{D}$ and a given distribution ν. This is made possible by a refined analysis of the `Successive-Rejects` strategy of Audibert et al. (2010). We finally provide lower bounds on the same average log-probability, also in terms of the same new information-theoretic quantities; these lower bounds are larger when the (natural) assumptions on the considered strategies are stronger. All these new upper and lower bounds generalize existing bounds based, e.g., on gaps between distributions. The content of this chapter is extracted from the conference paper

A. Barrier, A. Garivier, and G. Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 136–181. PMLR, 2023

Contents

5.1. Introduction

We consider a class $\mathcal{D}$ of distributions over $\mathbb{R}$ with finite first moments, which we refer to as the model $\mathcal{D}$. A K–armed bandit problem in $\mathcal{D}$ is a K–tuple $\underline{\nu} = (\nu_1, \dots, \nu_K)$ of distributions in $\mathcal{D}$. We denote by $(\mu_1, \dots, \mu_K)$ the K–tuple of their expectations. An agent sequentially interacts with $\underline{\nu}$: at each step $t \geq 1$, she selects an arm A_t and receives a reward Y_t drawn from the distribution ν_{A_t}. This is the only feedback that she obtains.

While regret minimization has been vastly studied (see Lattimore and Szepesvári, 2020), another relevant objective is *best-arm identification*, that is, identifying the distribution with highest expectation. In the fixed-confidence setting, this identification is performed under the constraint that a given confidence level $1 - \delta$ is respected, while minimizing the expected number of pulls of the arms (the expected sample complexity). This setting is fairly well understood (see Lattimore and Szepesvári, 2020, Chapter 33 for a review). A turning point in this literature was achieved by Garivier and Kaufmann (2016), who provided matching upper and lower bounds on the expected number of pulls of the arms in the case of canonical one-parameter exponential families. Since then, improvements have been made in several directions, including for example non-asymptotic bounds (Degenne et al., 2019) and the problem of ε–best-arm identification (Garivier and Kaufmann, 2021). The first generalization to non-parametric models in this fixed-confidence setting was achieved by Jourdan and Degenne (2023), who worked in a concurrent and independent manner from us. Their upper and lower bounds differ by a multiplicative factor of 2 (only).

Fixed-budget best-arm identification. The *fixed-budget setting* seems to be much less understood. Therein, the total number T of pulls of the arms is fixed. After these T pulls, a strategy must issue a recommendation $\hat{a}_T$. Assuming that $\underline{\nu}$ contains a unique optimal distribution $\nu^\star$ of index $a^\star(\underline{\nu})$, one aims at minimizing $\mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu}))$. We are interested in (upper and lower) bounds that hold for all problems $\underline{\nu}$ in $\mathcal{D}$, possibly under the restriction that they only contain a unique optimal arm. It may be straightforwardly seen that the probability of error can decay exponentially fast —for instance, by uniformly exploring the arms (pulling each of them about T/K times) and recommending the one with the largest empirical average. This is why the literature (see, for instance, Audibert et al., 2010 and Lattimore and Szepesvári, 2020, Chapter 33) focuses on lower and upper bound functions $\ell \leq U < 0$ of the typical form: *for all bandit problems* $\underline{\nu}$ in $\mathcal{D}$, with a unique optimal arm,

$$\ell(\underline{\nu}) \leq \liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq \limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq U(\underline{\nu}) < 0\,,$$

or, put differently, $$\exp\Big(\ell(\underline{\nu})\,T(1+o(1))\Big) \leq \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq \exp\Big(U(\underline{\nu})\,T(1+o(1))\Big)\,.$$

This problem is generally considered more difficult than the fixed-confidence setting (see, e.g., Lattimore and Szepesvári, 2020, Chapter 33 and Jourdan and Degenne, 2023, Section 6), and even for parametric models like canonical one-parameter exponential models, no strategy with matching upper and lower bounds (i.e., no optimal strategy) is known so far.

Earlier approaches. So far, four main approaches were considered for the problem of best-arm identification with a fixed budget. *First*, the early approach by Audibert et al. (2010) relies on gaps: we define the gap Δ_a of arm a as the difference $\mu^\star - \mu_a$ between the largest expectation $\mu^\star$ in $\underline{\nu}$ and the expectation of the distribution ν_a. They introduce a `Successive-Rejects` strategy and provide gap-based upper bounds for sub-Gaussian models, based on Hoeffding's inequality. They however propose a lower bound only in the case of a Bernoulli model, not for larger, non-parametric, models. This lower bound was further discussed by Carpentier and Locatelli (2016), in a minimax sense. *A second series of approaches* (see, e.g., Kaufmann et al., 2016) focused on Gaussian bandits with fixed variances, but their results do not seem to be easily generalized to other models as they rely on specific properties (even stronger than the symmetry of the Kullback-Leibler divergence, namely, that in this model, the Kullback-Leibler divergence only depends on the gap between the expectations of the distributions). *A third approach*, led by Russo (2016, 2020), considered canonical one-parameter exponential families, but for a different target probability. Namely, a Bayesian setting is considered and the quality of a strategy is measured as the posterior probability of identifying the best arm. An optimal non-gap-based complexity is exhibited, together with optimal strategies matching this complexity. However, Komiyama (2022) argues that such an approach is specific to the Bayesian case and is not suited to the frequentist case that we consider. *A fourth approach* is to focus on the case of $K = 2$ arms, see, e.g., Kaufmann et al. (2016). The non-parametric bounds obtained therein do not enjoy any obvious generalization to the case of $K \geq 3$ arms beyond the one stated in Theorem 5.14 and criticized in Section 5.2.3 for only involving pairwise comparisons with the best arm. By considering very specific models, Kato et al. (2022) constructed a strategy that is optimal (only) in the regime where the gap between the 2 arms is small —yet, this gap-based approach does not, by nature, go in the direction of non-parametric bounds.

We will provide more details concerning some of these approaches while presenting and discussing our main results, in Section 5.2.2; see also Section 5.6.

Outline and contributions. In this chapter, we focus our attention on instance-dependent upper and lower bounds, holding for all problems of general models $\mathcal{D}$, including non-parametric models, and valid for any number K of arms. Put differently, we target a high degree of generality. While admittedly not exhibiting matching upper and lower bounds, we show that the same (new) information-theoretic quantities $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{>}_{\inf}$ are at stake in these upper and lower bounds. These information-theoretic quantities are defined, in Section 5.2.1, as infima of Kullback-Leibler divergences and provide a quantification of the difficulty of the identification in terms of the geometry of information of the problem. We also present in Section 5.2.2 an overview of our results, which we carefully compare to existing bounds (restated therein, occasionally with some improvements). We state upper bounds in Section 5.3 and to do so, we provide an improved analysis of the classical `Successive-Rejects` strategy, not relying on gaps through Hoeffding's lemma. Section 5.4 exhibits several possible lower bounds, which are inversely larger to the strength of the assumptions made on the strategies. These lower bounds generalize known lower bounds in the literature, like the lower bound for Bernoulli models by Audibert et al. (2010), but hold for arbitrary models. They share some similar flavor with the lower bounds by Lai and Robbins (1985) and Burnetas and Katehakis (1996) for the cumulative regret.

Notation. For a given strategy facing a bandit problem $\underline{\mu}$, let $N_a(t)$ and $\hat{\mu}_a(t)$ denote the number of pulls and the empirical mean[1] of arm a at step t:

$$N_a(t) \overset{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = a\} \qquad \text{and} \qquad \hat{\mu}_a(t) \overset{\text{def}}{=} \frac{1}{N_a(t)} \sum_{s\in[t]} Y_s \mathbb{I}\{A_s = a\} \ .$$

[1] As strategies initially observe each arm once, $\hat{\mu}_a(t)$ is well-defined for $t \geq K$.

Without loss of generality (see the paragraph on optional skipping page 37), we assume that the observation at time step t is $Y_t = X_{A_t,N_{A_t}(t)}$, where $(X_{a,n})_{a\in[K],n\geq 1}$ are independent random variables such that $X_{a,n} \sim \mathcal{N}(\mu_{A_t}, 1)$ for all $a \in [K]$ and $n \geq 1$. As a consequence, we notably get

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{n=1}^{N_a(t)} X_{a,n} \overset{\text{def}}{=} \hat{\mu}_{a,N_a(t)} \,.$$

5.2. Overview of the Results and more Extended Literature Review

5.2.1. The Key new Quantities: $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{\leq}_{\inf}$, as well as $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$

In this chapter, we only consider models $\mathcal{D}$ whose distributions all admit an expectation. We denote by $\mathrm{E}(\zeta)$ the expectation of a distribution $\zeta \in \mathcal{D}$. For a distribution $\nu \in \mathcal{D}$ and a real number $x \in \mathbb{R}$, we then introduce

$$\mathcal{L}^{<}_{\inf}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) < x\}$$
$$\text{and} \qquad \mathcal{L}^{\leq}_{\inf}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) \leq x\}\,,$$

where KL denotes the Kullback-Leibler divergence and with the usual convention that the infimum of an empty set equals $+\infty$. Symmetrically, by considering rather distributions ζ with expectations larger than x, we define

$$\mathcal{L}^{>}_{\inf}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) > x\}$$
$$\text{and} \qquad \mathcal{L}^{\geq}_{\inf}(x,\nu) = \inf\{\mathrm{KL}(\zeta,\nu) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) \geq x\}\,.$$

We state some general properties on these quantities in Section 5.5.1 —among others, that $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{\leq}_{\inf}$, as well as $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$, are almost identical for the model $\mathcal{P}[0,1]$. The same holds for canonical one-parameter exponential models, as discussed in Section 5.5.3 (see page 151). In our results, lower bounds will be typically expressed with $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{>}_{\inf}$ quantities, while upper bounds will rely on $\mathcal{L}^{\leq}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$ quantities.

Remark. The key quantities for the non-parametric study of best-arm identification with fixed-confidence by Jourdan and Degenne (2023) are defined based on Kullback-Leibler divergences with arguments in reverse order, namely,

$$\mathcal{K}^{-}_{\inf}(\nu,x) = \inf\{\mathrm{KL}(\nu,\zeta) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) < x\} = \mathcal{K}_{\inf}(\nu,x)$$
$$\text{and} \qquad \mathcal{K}^{+}_{\inf}(\nu,x) = \inf\{\mathrm{KL}(\nu,\zeta) : \zeta \in \mathcal{D} \text{ s.t. } \mathrm{E}(\zeta) > x\}\,,$$

where the first quantity was referred to as simply $\mathcal{K}_{\inf}(\nu,x)$ by Honda and Takemura (2015) in the regret-minimization literature (see also Section 5.5.3 and Garivier et al., 2022). Optimal bounds for regret minimization only depend on $\mathcal{K}_{\inf}(\nu,x)$.

The introduction of $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$, and $\mathcal{L}^{\geq}_{\inf}$ are motivated by the fact that, in fixed-budget best-arm identification, the arguments in the KL are in reverse order compared to the fixed-confidence setting, see Section 5.2.2 (and also Section 2.3). Except for very specific models (e.g., the model $\mathcal{D}_{\sigma^2}$ of Gaussian distributions with a fixed variance $\sigma^2 > 0$), the Kullback-Leibler divergence is not symmetric, i.e., $\mathrm{KL}(\zeta,\nu)$ and $\mathrm{KL}(\nu,\zeta)$ differ in general. Specific best-arm-identification results were obtained by Kaufmann et al. (2016) for the model $\mathcal{D}_{\sigma^2}$, based on the Bretagnolle-Huber inequality (Bretagnolle and Huber, 1979); they indicate that the sum of the inverse squared gaps would be driving both the lower bound and upper bound functions ℓ and U. However, a close look at the proof reveals that they heavily rely on a property even stronger than the symmetry of KL for this model: details and discussions on this matter are provided in Section 5.6.2. In particular, generalizations beyond the Gaussian case appear to be infeasible.

5.2.2. Overview of the Results

The chapter provides new and more general (possibly non-parametric) bounds on the misidentification errors based on the information-theoretic quantities introduced above. In particular, we consider a version of Chernoff information defined, for ν, ν' in $\mathcal{D}$ with $\mathrm{E}(\nu') < \mathrm{E}(\nu)$, as

$$\mathcal{L}(\nu', \nu) = \inf_{x \in [\mathrm{E}(\nu'), \mathrm{E}(\nu)]} \left\{ \mathcal{L}^{\geq}_{\inf}(x, \nu') + \mathcal{L}^{\leq}_{\inf}(x, \nu) \right\}. \tag{5.1}$$

Given a bandit problem $\underline{\nu}$ with a unique optimal distribution denoted by $\nu^\star$, we may rank the arms a in non-decreasing order of $\mathcal{L}(\nu_a, \nu^\star)$, i.e., consider the permutation σ such that

$$0 = \mathcal{L}(\nu_{\sigma_1}, \nu^\star) < \mathcal{L}(\nu_{\sigma_2}, \nu^\star) \leq \ldots \leq \mathcal{L}(\nu_{\sigma_{K-1}}, \nu^\star) \leq \mathcal{L}(\nu_{\sigma_K}, \nu^\star). \tag{5.2}$$

Upper bound. *Our first main result* (Corollary 5.4 together with Lemma 5.5) considers models $\mathcal{D}$ like $\mathcal{D} = \mathcal{P}[0,1]$, the set of all probability distributions over $[0,1]$, or $\mathcal{D} = \mathcal{D}_{\exp}$, any canonical one-parameter exponential family. We study the `Successive-Rejects` strategy, introduced by Audibert et al. (2010), for which arms are rejected one by one at the end of phases of uniform exploration, and state that this strategy is such that for all bandit problems $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal arm,

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{\overline{\log} K} \min_{2 \leq k \leq K} \frac{\mathcal{L}(\nu_{\sigma_k}, \nu^\star)}{k}, \tag{5.3}$$

where $\overline{\log} K$ is defined in (5.18) and is of order $\log K$. The key for this result (Lemma 5.2, of independent interest) is a grid-based application of the Cramér-Chernoff bound to control $\mathbb{P}(\overline{X}_N \leq \overline{Y}_N)$, where $\overline{X}_N$ and $\overline{Y}_N$ are averages of two independent N–samples. This approach can be used to analyze similar algorithms, like `Sequential-Halving` (Karnin et al., 2013).

Lower bounds. The corresponding lower bounds are stated rather in terms of $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{>}_{\inf}$ quantities, but Sections 5.5.1 and 5.5.3 (page 151) explain that those quantities are almost the same than $\mathcal{L}^{\leq}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$ for regular models like $\mathcal{P}[0,1]$ and exponential models (for those models, $\mathcal{L}(\nu', \nu)$ could be alternatively defined with $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{>}_{\inf}$ instead of $\mathcal{L}^{\leq}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$, except in a single pathological case of the $\mathcal{P}[0,1]$ model). We actually state several lower bounds in Section 5.4, that are larger as the assumptions on the strategies considered are more restrictive; as usual, there is a trade-off between the strength of a lower bound and its generality. However, all assumptions considered remain rather mild and are empirically satisfied by `Successive-Rejects`-type strategies: for instance, Definition 5.8 restricts the attention to strategies such that for all bandit problems, the arm associated with the smallest expectation is pulled less than a fraction $\frac{1}{K}$ of the time. Out of all lower bounds exhibited, *our second main result* (Theorem 5.13) holds, as indicated, under mild assumptions on the model and sequences of strategies considered, and reads: for all bandit problems $\underline{\nu}$ with no two same expectations,

$$\liminf_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\min_{2 \leq k \leq K} \inf_{x \in [\mu_{(k)}, \mu_{(k-1)})} \left\{ \frac{\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})}{k-1} + \frac{\mathcal{L}^{<}_{\inf}(x, \nu^\star)}{k} \right\}, \tag{5.4}$$

where $\mu_{(1)} > \mu_{(2)} > \mu_{(3)} > \cdots > \mu_{(K)}$ and where $\nu_{(a)}$ denotes the distribution with expectation $\mu_{(a)}$. Here, we considered the notation (k) for order statistics in reverse order.

This lower bound does not match the exhibited upper bound, as is further discussed in Section 5.2.4. Still, we argue that quantities defined as infima over x of $\mathcal{L}^{>}_{\inf}(x, \nu_{(k)}) + \mathcal{L}^{<}_{\inf}(x, \nu^\star)$ should measure how difficult a best-arm-identification problem is under a fixed budget. *This is the main insight of this chapter.*

5.2.3. Re-Derivation of Existing Bounds

We now survey the most important existing bounds and re-derive them from our general bounds. These existing bounds all hold only for sub-Gaussian models and for exponential models when $K \geq 3$, while a non-parametric bound was only available in the case of $K = 2$ arms.

To do so, we will sometimes consider the following weaker version of the lower bound (5.4), obtained by picking $x = \mu_{(k)}$:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2\leq k\leq K} \frac{\mathcal{L}^<_{\inf}(\mu_{(k)}, \nu^\star)}{k} \,. \tag{5.5}$$

Comparison to the gap-based approaches. Audibert et al. (2010) propose an analysis of the `Successive-Rejects` strategy based on Hoeffding's inequality, stating that for all bandit problems in $\mathcal{P}[0,1]$ with a unique optimal arm,

$$\limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq - \frac{1}{\overline{\log} K} \min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k} \,, \tag{5.6}$$

where we recall the definition of the gaps $\Delta_{(k)} = \mu^\star - \mu_{(k)}$. This bound is a consequence of (Corollary 5.4, a slightly more general form of) the bound (5.3), given Pinsker's inequality (5.26):

$$\mathcal{L}(\nu_{(k)}, \nu^\star) \geq \inf_{x\in[\mu_{(k)},\mu^\star]} \Big\{ 2(x-\mu_{(k)})^2 + 2(x-\mu^\star)^2 \Big\} = (\mu^\star - \mu_{(k)})^2 = \Delta_{(k)}^2 \,. \tag{5.7}$$

We remark that the bound (5.6) and the lower bound on $\mathcal{L}(\nu_{(k)}, \nu^\star)$ may actually be extended to the model of σ^2–sub-Gaussian distributions, up to considering factors $\frac{1}{4\sigma^2}$. We do not discuss the `UCB-E` algorithm of Audibert et al. (2010), as its performance and analysis crucially depend on a tuning parameter set with some knowledge of the gaps.

Audibert et al. (2010) also propose a carefully constructed lower bound for the model $\mathcal{B}_{[p,\,1-p]} = \{\mathrm{Ber}(x) : x \in [p, 1-p]\}$ of Bernoulli distributions $\mathrm{Ber}(x)$ with parameters x in $[p, 1-p]$ for some $p \in (0, \frac{1}{2})$. A key inequality in their proof follows from the Kullback-Leibler – χ^2-divergence bound:

$$\forall x, y \in [p, 1-p], \qquad \mathrm{KL}(\mathrm{Ber}(x), \mathrm{Ber}(y)) \leq \frac{(x-y)^2}{2p(1-p)} \,.$$

Their construction may actually be generalized to models $\mathcal{D}$ with $C_{\mathcal{D}} > 0$ such that for all ν, ν' in $\mathcal{D}$, one has $\mathrm{KL}(\nu, \nu') \leq C_{\mathcal{D}} (\mathrm{E}(\nu) - \mathrm{E}(\nu'))^2$. This is a property that clearly holds for some exponential families: on top of the restricted Bernoulli model discussed above, for which

$$C_{\mathcal{B}_{[p,\,1-p]}} = \frac{1}{(2p(1-p))} \,,$$

we may cite the model $\mathcal{D}_{\sigma^2}$ of Gaussian distributions with variance σ^2, for which $C_{\mathcal{D}_{\sigma^2}} = 1/(2\sigma^2)$. For models enjoying the existence of such a constant $C_{\mathcal{D}}$, (a straightforward modification of) the analysis by Audibert et al. (2010) entails that for any $\underline{\nu}$ in $\mathcal{D}$,

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -5\, C_{\mathcal{D}} \min_{2\leq k\leq K} \frac{\Delta_{(k)}^2}{k} \,. \tag{5.8}$$

As by the very assumption on the model, $\mathcal{L}^<_{\inf}(\mu_{(k)}, \nu^\star) \leq C_{\mathcal{D}}\, \Delta_{(k)}^2$, the lower bound (5.5) implies the stated lower bound (5.8), with an improved constant factor.

The lower bound (5.8) and the upper bound (5.6) differ in particular by a factor proportional to $\overline{\log} K$. Carpentier and Locatelli (2016) discuss this gap in the case of the Bernoulli model $\mathcal{B}_{[1/4,\,3/4]}$ and improve the lower bound (5.8) by a factor of $\log K$, but not simultaneously for all bandit problems $\underline{\nu}$ (as we aim for); they obtain the improvement just for one bandit problem $\underline{\nu}$. Their lower bound result (formally stated and discussed in Section 5.6.1) is therefore of a totally different nature. More results on how and when given lower bounds with a given complexity measure may, or may not, be improved were stated by Komiyama et al. (2022).

Discussion on the non-parametric bound for $K = 2$ arms of Kaufmann et al. (2016). It turns out that the existing literature for the fixed-budget setting offered so far a non-parametric bound, in the case of $K = 2$ arms. Namely, in a general, possibly non-parametric model $\mathcal{D}$, Kaufmann et al. (2016, Theorem 12) stated a lower bound for all 2–armed bandit problems $\underline{\nu} = (\nu_1, \nu_2)$:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \inf_{\substack{\underline{\lambda} \text{ in } \mathcal{D}: \\ \mathrm{E}(\lambda_{a^\star(\underline{\nu})}) < \mathrm{E}(\lambda_{w_\star(\underline{\nu})})}} \max\Big\{ \mathrm{KL}(\lambda_{w_\star(\underline{\nu})}, \nu_{w_\star(\underline{\nu})}),\, \mathrm{KL}(\lambda_{a^\star(\underline{\nu})}, \nu_{a^\star(\underline{\nu})}) \Big\}, \quad (5.9)$$

where $w_\star(\underline{\nu})$ denotes the sub-optimal arm in $\underline{\nu}$ and where the infimum is over all alternative bandit problems (λ_1, λ_2) in $\mathcal{D}$ with a different best arm thant $\underline{\nu}$. We note (see the proof of Theorem 5.14) that we may actually rewrite this lower bound in a more readable way, in terms of $\mathcal{L}^<_{\inf}$ and $\mathcal{L}^>_{\inf}$ quantities, illustrating once again that these quantities are key in measuring the complexity of best-arm identification under a fixed budget:

$$\inf_{\substack{\underline{\lambda} \text{ in } \mathcal{D}: \\ \mathrm{E}(\lambda_{a^\star(\underline{\nu})}) < \mathrm{E}(\lambda_{w_\star(\underline{\nu})})}} \max\Big\{ \mathrm{KL}(\lambda_{w_\star(\underline{\nu})}, \nu_{w_\star(\underline{\nu})}),\, \mathrm{KL}(\lambda_{a^\star(\underline{\nu})}, \nu_{a^\star(\underline{\nu})}) \Big\}$$
$$= \inf_{x \in [\mu_{w_\star(\underline{\nu})}, \mu^\star]} \Big\{ \max\Big\{ \mathcal{L}^>_{\inf}(x, \nu_{w_\star(\underline{\nu})}),\, \mathcal{L}^<_{\inf}(x, \nu^\star) \Big\} \Big\}. \quad (5.10)$$

The proof technique of Kaufmann et al. (2016) may be applied in a pairwise fashion to generalize the lower bound (5.10) for 2 arms into a lower bound for $K \geq 2$ arms, stated in Theorem 5.14: for all $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal arm,

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{k \neq a^\star(\underline{\nu})} \inf_{x \in [\mu_k, \mu^\star]} \Big\{ \max\big\{ \mathcal{L}^>_{\inf}(x, \nu_k),\, \mathcal{L}^<_{\inf}(x, \nu^\star) \big\} \Big\}. \quad (5.11)$$

We do not claim that (5.11) is a deep and interesting bound, as it only involves pairwise comparisons with the best arm. In particular, we lack divisions by the ranks of the arms, as in (5.4). This is why we had not stated the result (5.11) of Theorem 5.14 in Section 5.2.2 and mention it only here.

That being said, given that the infima in (5.4) are over more restricted ranges than in (5.11), we can see no obvious ranking between the two bounds, which rather look incomparable.

Bounds for $K = 2$ arms and exponential families, cf. comments after Theorem 12 of Kaufmann et al. (2016). We denote by $\mathcal{D}_{\exp}$ the model corresponding to a canonical one-parameter exponential family with expectations defined on an open interval $\mathcal{M}$ (see page 151 in Section 5.5.3 for a reminder on this matter). For such a model, we denote by d the mean-parameterized Kullback-Leibler divergence. By continuity of d, we have that for all ν in $\mathcal{D}_{\exp}$ and for all $x \in \mathcal{M}$,

$$\forall x \leq \mathrm{E}(\nu), \qquad \mathcal{L}^<_{\inf}(x, \nu) = \mathcal{L}^\leq_{\inf}(x, \nu) = d(x, \mathrm{E}(\nu)), \quad (5.12)$$

$$\text{and} \qquad \forall x \geq \mathrm{E}(\nu), \qquad \mathcal{L}^>_{\inf}(x, \nu) = \mathcal{L}^\geq_{\inf}(x, \nu) = d(x, \mathrm{E}(\nu)). \quad (5.13)$$

Note that all bounds stated in Section 5.2.2 then admit simple reformulations in terms of d. The Chernoff-information-type quantity $\mathcal{L}$ introduced in (5.1) may also be mean-parameterized as follows: for $\mu' < \mu$,

$$L(\mu', \mu) = \min_{x \in [\mu', \mu]} \big\{ d(x, \mu') + d(x, \mu) \big\}. \quad (5.14)$$

We now explain why we called L (and therefore $\mathcal{L}$) a version of Chernoff information. The original definition of the Chernoff information $D(\mu', \mu)$ is the value $d(y, \mu)$ for $y \in [\mu', \mu]$ such that $d(y, \mu') = d(y, \mu)$. As mentioned in the comments after Theorem 12 of Kaufmann et al. (2016), D is the quantity at stake in (5.10) for a canonical one-parameter exponential family: given that $d(\,\cdot\,, \mu')$ and $d(\,\cdot\,, \mu)$ are respectively increasing and decreasing on $[\mu', \mu]$,

$$\min_{x \in [\mu', \mu]} \max\{d(x, \mu'),\, d(x, \mu)\} = D(\mu', \mu)\,.$$

Therefore, $D(\mu', \mu) \leq L(\mu', \mu) \leq 2\, D(\mu', \mu)$, which shows that L is related to D, as claimed.

Example 5.1. We state the lower bound (5.5) and the upper bound (5.3) for the model $\mathcal{B}_{[p,\,1-p]}$ of Bernoulli distributions with parameters in $[p, 1-p]$, where $p \in (0, \frac{1}{2})$. We denote by

$$\mathrm{kl}(x, y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}\,, \qquad \text{where} \qquad x, y \in [p, 1-p]$$

the mean-parameterized Kullback-Leibler divergence of this model. We consider a generic bandit problem $\underline{\nu} = (\mathrm{Ber}(p_1), \dots, \mathrm{Ber}(p_K))$. We rank the parameters as in (5.4), i.e., introduce the notation $p^\star = p_{(1)} > p_{(2)} > \ldots > p_{(K)}$. Then, after noticing (see Lemma 5.19 in Section 5.5.3) that this ranking is the same as the one considered in (5.2), the upper bound (5.3) rewrites as

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq -\frac{1}{\overline{\log}\, K} \min_{2 \leq k \leq K} \frac{\min\limits_{x \in [p_{(k)}, p^\star]} \Big\{ \mathrm{kl}(x, p_{(k)}) + \mathrm{kl}(x, p^\star) \Big\}}{k}\,,$$

while the lower bound (5.5) rewrites as

$$\liminf_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\min_{2 \leq k \leq K} \frac{\mathrm{kl}(p_{(k)}, p^\star)}{k}\,.$$

They should be compared to the upper (5.6) and lower (5.8) bounds of Audibert et al. (2010), respectively.

5.2.4. Discussion of the (Lack of) Optimality of the new Bounds Exhibited

The lower bound (5.4) does not match the upper bound (5.3) because of two aspects. First, the infima in (5.4) are only taken on restricted ranges $[\mu_{(k)}, \mu_{(k-1)})$ and not on the entire intervals $[\mu_{(k)}, \mu^\star]$ as in (5.3). Second, the upper bound (5.3) involves a $1/\overline{\log} K$ factor, while the lower bound (5.4) does not. A similar $1/\overline{\log} K$ factor was missing between the upper (5.6) and lower (5.8) bounds of Audibert et al. (2010) for Bernoulli models, together with a numerical factor of $5\, C_{\mathcal{B}_{[p,\,1-p]}}$. The non-parametric bounds exhibited in this chapter mainly generalize and extend the known parametric bounds but do not refine the latter in the sense that gaps between upper and lower bounds would be closed.

That being said, we would like to illustrate below one specific example to which extent the gap-based bounds can be looser.

Example of an extreme improvement: distributions with separated supports. For general non-parametric models, gaps are not enough at all to measure complexity as we may well have a finite gap between two distributions ν_1 and ν_2 with $\mu_1 > \mu_2$, but $\mathcal{L}(\nu_2, \nu_1) = +\infty$. This holds, for instance, as soon as ν_1 and ν_2 have closed supports separated by a threshold x_0 (see Figure 5.1), i.e., the closed supports of ν_1 and ν_2 are included in $(-\infty, x_0)$ and $(x_0, +\infty)$, respectively. Indeed,

by mimicking the beginning of the proof of Lemma 5.16, it may be seen that $\mathcal{L}_{\inf}^{\leq}(x, \nu_1) = +\infty$ for $x \leq x_0$ and $\mathcal{L}_{\inf}^{\geq}(x, \nu_2) = +\infty$ if $x \geq x_0$, so that in all cases, the sum $\mathcal{L}_{\inf}^{\geq}(x, \nu_2) + \mathcal{L}_{\inf}^{\leq}(x, \nu_1)$ equals $+\infty$, and thus, $\mathcal{L}(\nu_2, \nu_1) = +\infty$. In our bounds, e.g., the upper bound (5.3), the pair of distributions ν_1, ν_2 will therefore not contribute —as intuition commands: these two distributions are easy to distinguish—, while it does contribute to the earlier gap-based bounds.

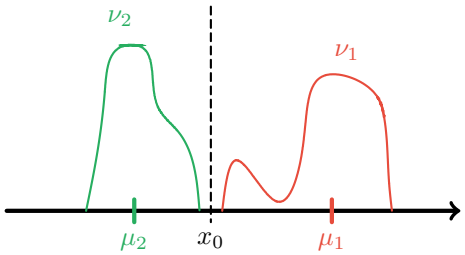


Figure 5.1: When the two distributions have separated supports, there is no confusion to identify which of both has the highest mean.

5.3. Upper Bound for the `Successive-Rejects` Strategy, with an Improved Analysis

We consider the `Successive-Rejects` strategy introduced by Audibert et al. (2010), for K arms and a budget T. The strategy works in phases, and the lengths of the phases are set beforehand; they are denoted by $\ell_1, \ldots, \ell_{K-1} \geq 1$ and satisfy $\ell_1 + \ldots + \ell_{K-1} = T$. The strategy maintains a list of candidate arms, starting with all arms, i.e., $S_0 = [K]$. At the end of each phase $r \in \{1, \ldots, K-1\}$, it drops an arm to get S_r, while during phase r, it operates with the $K - r + 1$ arms in S_{r-1}.

More precisely, during phase $r \in \{1, \ldots, K-1\}$, the strategy draws $\left\lfloor \frac{\ell_r}{K-r+1} \right\rfloor$ times each arm in S_{r-1} (and does not use the few remaining time steps, if there are some). At the end of each phase r, the strategy computes the empirical averages $\overline{X}_a^r$ of the payoffs obtained by each arm $a \in S_{r-1}$ since the beginning; i.e., $\overline{X}_a^r$ is an average over

$$N_r = \left\lfloor \frac{\ell_1}{K} \right\rfloor + \ldots + \left\lfloor \frac{\ell_r}{K-r+1} \right\rfloor$$

i.i.d. realizations of ν_a. It then drops the arm a_r with smallest empirical average (ties broken arbitrarily). This description is summarized in Algorithm 16.

5.3.1. General Analysis

The key quantities for the general analysis will be the logarithmic moment-generating function ϕ_ν of a distribution $\nu \in \mathcal{D}$, and its Fenchel-Legendre transform $\phi_\nu^\star$:

$$\forall \lambda \in \mathbb{R}, \quad \phi_\nu(\lambda) = \log \int_{\mathbb{R}} \mathrm{e}^{\lambda x} \, \mathrm{d}\nu(x) \qquad \text{and} \qquad \forall x \in \mathbb{R}, \quad \phi_\nu^\star(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \phi_\nu(\lambda)\} . \tag{5.15}$$

Based on them, we can now define, for all $\nu, \nu' \in \mathcal{D}$ with $\mathrm{E}(\nu') < \mathrm{E}(\nu)$,

$$\Phi(\nu', \nu) \stackrel{\text{def}}{=} \inf_{x \in [\mathrm{E}(\nu'), \mathrm{E}(\nu)]} \{\phi_{\nu'}^\star(x) + \phi_\nu^\star(x)\} .$$

The following lemma shows that Φ plays a significant role in bounding the probability that two sample averages are in reverse order compared to the expectations of the underlying distributions. It supersedes the use of Hoeffding's inequality in Audibert et al. (2010).

Algorithm 16: `Successive-Rejects` algorithm

Input: budget parameter T
phase lengths $(\ell_r)_{r\in[K-1]}$ such that $\sum_{r\in[K-1]} \ell_r = T$
Output: estimated best arm $\hat{a}_T$

$t \leftarrow 0$
$S_0 \leftarrow [K]$
for *each round* $r \in [K-1]$ **do**
  Observe each arm $\left\lfloor \frac{\ell_r}{K-r+1} \right\rfloor$ times
  Increase t by ℓ_r
  Choose $a_r \in \operatorname*{argmin}_{a\in S_{r-1}} \hat{\mu}_a(t)$
  $S_r \leftarrow S_{r-1} \setminus \{a_r\}$
Define $\hat{a}_T$ as the unique element of S_{K-1}

Lemma 5.2. *Fix ν and ν' in $\mathcal{D}$, with respective expectations $\mu = \mathrm{E}(\nu) > \mu' = \mathrm{E}(\nu')$. For all $N \geq 1$, let $\overline{X}_N$ and $\overline{Y}_N$ be the averages of N–samples with respective distributions ν and ν'. Then,*

$$\limsup_{N\to+\infty} \frac{1}{N} \log \mathbb{P}(\overline{X}_N \leq \overline{Y}_N) \leq - \inf_{x\in[\mu',\mu]} \{\phi^\star_{\nu'}(x) + \phi^\star_\nu(x)\} \stackrel{\text{def}}{=} -\Phi(\nu',\nu)\,.$$

Proof. The proof consists of two parts. We first show that for any finite grid $\mathcal{G} = \{g_2, \ldots, g_{G-1}\}$ in (μ', μ), to which we add the points $g_1 = \mu'$ and $g_G = \mu$, we have

$$\limsup_{N\to+\infty} \frac{1}{N} \log \mathbb{P}(\overline{X}_N \leq \overline{Y}_N) \leq - \min\Big\{\phi^\star_\nu(\mu'),\ \min_{2\leq j\leq G-1} \{\phi^\star_{\nu'}(g_{j-1}) + \phi^\star_\nu(g_j)\},\ \phi^\star_{\nu'}(\mu)\Big\}\,. \quad (5.16)$$

Indeed, by identifying (when $\overline{X}_N$ and $\overline{Y}_N$ belong to $[\mu', \mu]$) in which interval $[g_{j-1}, g_j]$ lies $\overline{X}_N$, we note that

$$\{\overline{X}_N \leq \overline{Y}_N\} \quad \subseteq \quad \{\overline{X}_N \leq \mu'\} \ \cup\ \{\overline{Y}_N \geq \mu\} \ \cup \bigcup_{j=2}^{G-1} \{\overline{Y}_N \geq g_{j-1} \text{ and } \overline{X}_N \leq g_j\}\,.$$

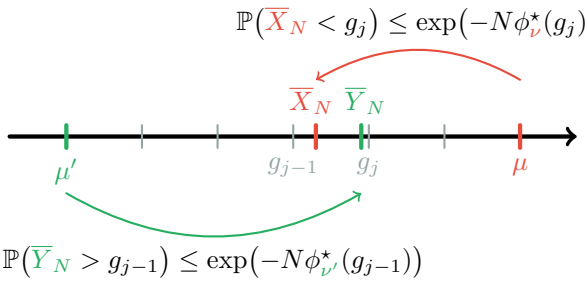


Figure 5.2: Control of the probability of $\{\overline{Y}_N \geq g_{j-1} \text{ and } \overline{X}_N \leq g_j\}$ using a Cramér-Chernoff bound.

First, by independence and by the Cramér-Chernoff inequalities (5.29) and (5.30),

$$\mathbb{P}(\overline{Y}_N \geq g_{j-1} \text{ and } \overline{X}_N \leq g_j) = \mathbb{P}(\overline{Y}_N \geq g_{j-1})\ \mathbb{P}(\overline{X}_N \leq g_j) \leq \exp\Big(-N\big(\phi^\star_{\nu'}(g_{j-1}) + \phi^\star_\nu(g_j)\big)\Big)\,.$$

Second, again by the Cramér-Chernoff inequalities,

$$\mathbb{P}(\overline{X}_N \leq \mu') \leq \exp(-N\,\phi_\nu^\star(\mu')) \qquad \text{and} \qquad \mathbb{P}(\overline{Y}_N \geq \mu) \leq \exp(-N\,\phi_{\nu'}^\star(\mu))\,.$$

By a union bound,

$$\mathbb{P}(\overline{X}_N \leq \overline{Y}_N) \leq \exp(-N\,\phi_\nu^\star(\mu')) + \exp(-N\,\phi_{\nu'}^\star(\mu)) + \sum_{j=2}^{G-1} \exp\Big(-N(\phi_{\nu'}^\star(g_{j-1}) + \phi_\nu^\star(g_j))\Big)\,.$$

The stated bound (5.16) follows by identifying the (finitely many) terms with the smallest rate in the exponent.

In the second part of the proof, we note that the bound (5.16) holds for any finite grid in (μ', μ), and we consider a sequence

$$\mathcal{G}^{(n)} = \Big\{g_2^{(n)}, \dots, g_{G_n-1}^{(n)}\Big\}$$

of such finite grids. In particular,

$$\limsup_{N\to+\infty} \frac{1}{N} \log \mathbb{P}(\overline{X}_N \leq \overline{Y}_N) \leq -\min\Big\{\phi_\nu^\star(\mu'),\ \max_{n\geq 1} S_n,\ \phi_{\nu'}^\star(\mu)\Big\}\,,$$
$$\text{where} \qquad S_n \overset{\text{def}}{=} \min_{2\leq j\leq G_n-1} \Big\{\phi_{\nu'}^\star(g_{j-1}^{(n)}) + \phi_\nu^\star(g_j^{(n)})\Big\}\,.$$

To obtain the claimed bound, given that (see the end of Section 5.5.2)

$$\phi_\nu^\star(\mu) = 0 = \phi_{\nu'}^\star(\mu')\,,$$

it suffices to show that

$$\max_{n\geq 1} S_n \geq \inf_{x\in[\mu',\mu]} \{\phi_{\nu'}^\star(x) + \phi_\nu^\star(x)\}\,.$$

To that end, we assume that the steps ε_n of the grids $\mathcal{G}^{(n)}$, which are defined as

$$\varepsilon_n \overset{\text{def}}{=} \max_{2\leq j\leq G_n} \Big|g_j^{(n)} - g_{j-1}^{(n)}\Big|\,,$$

vanish asymptotically, i.e., $\varepsilon_n \to 0$. For each grid $\mathcal{G}^{(n)}$, we denote by $x_n^\star \in (\mu', \mu)$ the argument of the minimum in the definition of S_n. As a consequence, for each $n \geq 1$,

$$S_n = \phi_{\nu'}^\star(x_n^\star - \varepsilon_n^\star) + \phi_\nu^\star(x_n^\star)\,,$$

for some $0 < \varepsilon_n^\star \leq \varepsilon_n$. The quantity $x_n^\star - \varepsilon_n^\star$ denotes the point in the grid that is right before $x_n^\star$, and it belongs to $[\mu', \mu)$. We note that we also have $\varepsilon_n^\star \to 0$. In the compact interval $[\mu', \mu]$, the Bolzano-Weierstrass theorem (see, e.g., Bartle and Sherbert, 2000, Section 3.4) ensures the existence of a converging sub-sequence: there exists $x_\infty^\star \in [\mu', \mu]$ and a sequence $(n_k)_{k\geq 1}$ of integers such that

$$x_{n_k}^\star \underset{k\to+\infty}{\longrightarrow} x_\infty^\star\,, \qquad \text{which also entails} \qquad x_{n_k}^\star - \varepsilon_{n_k}^\star \underset{k\to+\infty}{\longrightarrow} x_\infty^\star\,.$$

Now, the functions $\phi_\nu^\star$, respectively, $\phi_{\nu'}^\star$, are lower semi-continuous, as the suprema over $\lambda \in \mathbb{R}$ of the continuous functions $x \mapsto \lambda x - \varphi_\nu(\lambda)$, respectively, $x \mapsto \lambda x - \varphi_{\nu'}(\lambda)$. Therefore, by these lower semi-continuities,

$$\max_{n\geq 1} S_n \geq \liminf_{k\to+\infty} \phi_{\nu'}^\star(x_{n_k}^\star - \varepsilon_{n_k}^\star) + \phi_\nu^\star(x_{n_k}^\star) \geq \phi_{\nu'}^\star(x_\infty^\star) + \phi_\nu^\star(x_\infty^\star)$$
$$\geq \inf_{x\in[\mu',\mu]} \{\phi_{\nu'}^\star(x) + \phi_\nu^\star(x)\}\,.$$

This concludes the proof. □

General upper bound. The main performance upper bound is stated below in terms of Φ, that is, in terms of Fenchel-Legendre transforms of logarithmic moment-generating functions. Section 5.4.2 will later explain why and when the latter may be replaced by $\mathcal{L}^{\leq}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$ quantities, leading to a rewriting $\Phi = \mathcal{L}$ and to the bound claimed in (5.3).

Theorem 5.3. *Fix $K \geq 2$ and a model $\mathcal{D}$. Consider a sequence of* `Successive-Rejects` *strategies, indexed by T, such that $N_r/T \to \gamma_r > 0$ as $T \to +\infty$ for all $r \in \{1, \dots, K-1\}$. Let $\underline{\nu}$ be a bandit problem in $\mathcal{D}$ with a unique optimal arm and, for each $r \in \{1, \dots, K-1\}$, let $\mathcal{A}_r$ be a subset of arms of cardinality r that does not contain $a^\star(\underline{\nu})$. Then*

$$\limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq - \min_{1\leq r\leq K-1} \Big\{ \gamma_r \min_{k\in\mathcal{A}_r} \Phi(\nu_k, \nu^\star) \Big\}.$$

The proof mimics the analysis by Audibert et al. (2010), the main modification being the substitution of Hoeffding's inequality by the bound of Lemma 5.2.

Proof. We recall that for $r \in \{1, \dots, K-1\}$, we denoted by $N_r = \lfloor \ell_1/K \rfloor + \ldots + \lfloor \ell_r/(K-r+1) \rfloor$ the total number of times an arm still considered in phase r, i.e., belonging to S_{r-1}, was pulled in phases 1 to r. For each arm a, we denote by $\overline{Y}^r_a$ the average of a N_r–sample distributed according to ν_a. By optional skipping (see Doob, 1953, Chapter III, Theorem 5.2, p. 145, or Chow and Teicher, 1988, Section 5.3 for a more recent reference), we may assume, with no loss of generality, that for each $r \in \{1, \dots, K-1\}$,

$$\text{on the event } \{a \in S_{r-1}\}, \qquad \overline{X}^r_a = \overline{Y}^r_a. \tag{5.17}$$

We fix a bandit problem $\underline{\nu}$ with a unique optimal arm $a^\star(\underline{\nu})$. The `Successive-Rejects` strategy fails if (and only) if it rejects $a^\star(\underline{\nu})$ in one of the phases. This corresponds to the event

$$\{\hat{a}_T \neq a^\star(\underline{\nu})\} = \bigcup_{r=1}^{K-1} \{a_r = a^\star(\underline{\nu})\} \subseteq \bigcup_{r=1}^{K-1} \Big\{ a^\star(\underline{\nu}) \in S_{r-1} \text{ and } \forall k \in S_{r-1}, \ \overline{X}^r_{a^\star(\underline{\nu})} \leq \overline{X}^r_k \Big\},$$

where we recall that a_r is the arm removed at the end of phase r (we have an inclusion because ties are broken arbitrarily). By optional skipping (5.17),

$$\bigcup_{r=1}^{K-1} \Big\{ a^\star(\underline{\nu}) \in S_{r-1} \text{ and } \forall k \in S_{r-1}, \ \overline{X}^r_{a^\star(\underline{\nu})} \leq \overline{X}^r_k \Big\} = \bigcup_{r=1}^{K-1} \Big\{ a^\star(\underline{\nu}) \in S_{r-1} \text{ and } \forall k \in S_{r-1}, \ \overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k \Big\}.$$

Recall that the set S_{r-1} is a random set; dealing with it therefore requires some care. On the event of interest, S_{r-1} contains $K-r+1$ elements, among which $a^\star(\underline{\nu})$. The set $\mathcal{A}_r$ is of cardinality r and does not contain $a^\star(\underline{\nu})$. By the pigeonhole principle, S_{r-1} thus necessarily contains one arm in $\mathcal{A}_r$. As a consequence, for each phase $r \in \{1, \dots, K-1\}$,

$$\Big\{ a^\star(\underline{\nu}) \in S_{r-1} \text{ and } \forall k \in S_{r-1}, \ \overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k \Big\} \subseteq \bigcup_{k\in\mathcal{A}_r} \Big\{ \overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k \Big\}.$$

Summarizing the inclusions above, taking unions bounds, and upper bounding the obtained sum in a crude way, we proved so far

$$\mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq \sum_{r=1}^{K-1} \sum_{k\in\mathcal{A}_r} \mathbb{P}\Big(\overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k \Big) \leq K^2 \max_{1\leq r\leq K-1} \max_{k\in\mathcal{A}_r} \mathbb{P}\Big(\overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k \Big),$$

or equivalently,

$$\frac{1}{T}\log\mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq \frac{2}{T}\log K + \max_{1\leq r\leq K-1}\max_{k\in\mathcal{A}_r}\frac{1}{T}\log\mathbb{P}\Big(\overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k\Big)$$
$$= \frac{2}{T}\log K + \max_{1\leq r\leq K-1}\max_{k\in\mathcal{A}_r}\frac{N_r}{T}\frac{1}{N_r}\log\mathbb{P}\Big(\overline{Y}^r_{a^\star(\underline{\nu})} \leq \overline{Y}^r_k\Big)\,.$$

As $N_r/T \to \gamma_r > 0$ as $T \to +\infty$, we may apply Lemma 5.2, together with an exchange between the $\limsup$ and the maximum over a finite number of quantities. We obtain

$$\limsup_{T\to+\infty}\frac{1}{T}\log\mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu})) \leq \max_{1\leq r\leq K-1}\max_{k\in\mathcal{A}_r}\Big\{\gamma_r\Big(-\Phi(\nu_k,\nu^\star)\Big)\Big\}$$
$$= -\min_{1\leq r\leq K-1}\Big\{\gamma_r\min_{k\in\mathcal{A}_r}\Phi(\nu_k,\nu^\star)\Big\}\,.$$

This concludes the proof. □

With the phase lengths of Audibert et al. (2010). We conclude this subsection by stating the bound of Theorem 5.3 for the phase lengths suggested by Audibert et al. (2010), namely,

$$\ell_1 \overset{\text{def}}{=} \frac{T}{\overline{\log}\,K}\,, \qquad \text{and} \qquad \forall r\in\{2,\dots,K-1\}, \qquad \ell_r \overset{\text{def}}{=} \frac{T}{(K-r+2)\,\overline{\log}\,K}\,, \tag{5.18}$$

where we define

$$\overline{\log}\,K \overset{\text{def}}{=} \frac{1}{2} + \sum_{k=2}^{K}\frac{1}{k}\,,$$

The phase lengths in the example case of $K = 6$ arms are shown in Figure 5.3. Rather thant Φ, it

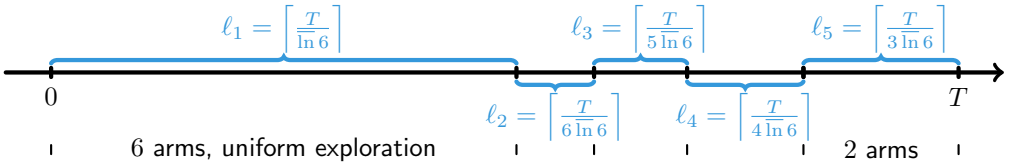


Figure 5.3: `Successive-Rejects` phase lengths of Audibert et al. (2010) for $K = 6$ arms.

is sometimes handy to rely on more readable quantities, this is why we will consider lower bounds $f(\nu_k,\nu^\star)$ on the $\Phi(\nu_k,\nu^\star)$ quantities. We may of course use $f = \Phi$ but f can also be, for instance, the squared gaps: in the case of the $\mathcal{P}[0,1]$ model, Hoeffding's inequality entails that

$$\phi^\star_\nu(x) \geq 2\big(x-\mathrm{E}(\nu)\big)^2\,, \qquad \text{so that} \qquad \Phi(\nu_k,\nu^\star) \geq \Delta_k^2 \overset{\text{def}}{=} f(\nu_k,\nu^\star)\,, \tag{5.19}$$

as explained in the proof below.

Remark. Such bounds hold more generally in models consisting of sub-Gaussian distributions. For ease of exposition, the path followed in Section 5.2.2 to show that $\Phi(\nu_k,\nu^\star) \geq \Delta_k^2$ was to first note that $\Phi = \mathcal{L}$ when $\mathcal{D} = \mathcal{P}[0,1]$ (see Lemma 5.5) and then use Pinsker's inequality (5.7). We provide below a slightly more direct but equivalent approach, based on Hoeffding's inequality.

Proof of (5.19). When $\nu \in \mathcal{P}[0,1]$, Hoeffding's inequality exactly states that

$$\forall\lambda\in\mathbb{R}, \quad \phi_\nu(\lambda) \leq \lambda\mathrm{E}(\nu) + \frac{\lambda^2}{8}\,,$$
$$\text{so that} \qquad \forall x\in\mathbb{R}, \quad \phi^\star_\nu(x) \geq \sup_{\lambda\in\mathbb{R}}\left\{\lambda(x-\mathrm{E}(\nu)) - \frac{\lambda^2}{8}\right\} = 2\big(x-\mathrm{E}(\nu)\big)^2\,.$$

This corresponds to the first part of (5.19).

For its second part, we consider a pair ν, ν' of distributions in $\mathcal{P}[0,1]$, we set any $x \in [\mathrm{E}(\nu'), \mathrm{E}(\nu)]$, and we apply twice the bound of the first part to get

$$\phi^\star_{\nu'}(x) + \phi^\star_{\nu}(x) \geq 2\big(x - \mathrm{E}(\nu')\big)^2 + 2\big(x - \mathrm{E}(\nu)\big)^2 .$$

From the definition of Φ, it follows that

$$\Phi(\nu', \nu) \geq \inf_{x \in [\mathrm{E}(\nu'), \mathrm{E}(\nu)]} \Big\{ 2\big(x - \mathrm{E}(\nu')\big)^2 + 2\big(x - \mathrm{E}(\nu)\big)^2 \Big\} = \big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)^2 .$$

This corresponds to the second part of (5.19). □

We now order the arms into $\sigma_1, \ldots, \sigma_K$ based on f, namely, we let $\sigma_1 = a^\star(\underline{\nu})$ and

$$0 = f(\nu_{\sigma_1}, \nu^\star) < f(\nu_{\sigma_2}, \nu^\star) \leq \ldots \leq f(\nu_{\sigma_{K-1}}, \nu^\star) \leq f(\nu_{\sigma_K}, \nu^\star) , \tag{5.20}$$

and we take $\mathcal{A}_r = \{\sigma_{K-r+1}, \ldots, \sigma_K\}$. We obtain the following corollary.

Corollary 5.4. *Fix $K \geq 2$, a model $\mathcal{D}$, and consider a lower bound f on Φ. The sequence of* `Successive-Rejects` *strategies based on the phase lengths* (5.18) *ensures, that for all bandit problems $\underline{\nu}$ in $\mathcal{D}$ with a unique optimal arm,*

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}\big(\hat{a}_T \neq a^\star(\underline{\nu})\big) \leq -\frac{1}{\overline{\log} K} \min_{2 \leq k \leq K} \frac{f(\nu_{\sigma_k}, \nu^\star)}{k} ,$$

where arms were reordered as in (5.20).

Proof. To apply Theorem 5.3, we need only to show that the phase lengths of (5.18) are such that N_r/T converges to a positive value, and to identify this limit value γ_r. As $N_1 = \lfloor \ell_1/K \rfloor$, where $\ell_1 = T/\overline{\log} K$, we immediately have $N_1/T \to \gamma_1 = 1/(K \overline{\log} K) > 0$. For $r \in \{2, \ldots, K-1\}$,

$$\begin{aligned} \frac{N_r}{T} = \sum_{p=1}^{r} \frac{1}{T} \left\lfloor \frac{\ell_p}{K} \right\rfloor &= \frac{1}{T} \left(\left\lfloor \frac{T}{K \overline{\log} K} \right\rfloor + \sum_{p=2}^{r} \left\lfloor \frac{T}{(K-p+1)(K-p+2) \overline{\log} K} \right\rfloor \right) \\ &\underset{T \to +\infty}{\longrightarrow} \gamma_r \overset{\text{def}}{=} \frac{1}{\overline{\log} K} \left(\frac{1}{K} + \sum_{p=2}^{r} \frac{1}{K-p+1} - \frac{1}{K-p+2} \right) = \frac{1}{(K-r+1) \overline{\log} K} . \end{aligned}$$

The bound of Theorem 5.3 reads:

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}\big(\hat{a}_T \neq a^\star(\underline{\nu})\big) \leq -\frac{1}{\overline{\log} K} \min_{1 \leq r \leq K-1} \left\{ \frac{1}{K-r+1} \min_{k \in \mathcal{A}_r} \Phi(\nu_k, \nu^\star) \right\} .$$

It implies, in terms of lower bounds $f(\nu_k, \nu^\star) \leq \Phi(\nu_k, \nu^\star)$,

$$\limsup_{T \to +\infty} \frac{1}{T} \log \mathbb{P}\big(\hat{a}_T \neq a^\star(\underline{\nu})\big) \leq -\frac{1}{\overline{\log} K} \min_{1 \leq r \leq K-1} \left\{ \frac{1}{K-r+1} \min_{k \in \mathcal{A}_r} f(\nu_k, \nu^\star) \right\} . \tag{5.21}$$

The permutation σ in (5.20) and the sets $\mathcal{A}_r = \{\sigma_{K-r+1}, \ldots, \sigma_K\}$ were exactly picked, for each $r \in \{1, \ldots, K-1\}$, to minimize

$$\min_{k \in \mathcal{B}_r} f(\nu_k, \nu^\star)$$

over sets $\mathcal{B}_r$ abiding by the indicated constraints: being of cardinal r and not containing the optimal arm $a^\star(\underline{\nu}) = \sigma_1$. We get

$$\min_{k \in \mathcal{A}_r} f(\nu_k, \nu^\star) = \min_{K-r+1 \leq k \leq K} f(\nu_{\sigma_k}, \nu^\star) = f(\nu_{\sigma_{K-r+1}}, \nu^\star) ,$$

which, together with (5.21), yields the stated bound, up to replacing $K-r+1$ with $r \in \{1,\dots,K-1\}$ by $k \in \{2,\dots,K\}$:

$$-\frac{1}{\overline{\log} K} \min_{1\le r\le K-1} \left\{ \frac{1}{K-r+1} f\left(\nu_{\sigma_{K-r+1}}, \nu^\star\right) \right\} = -\frac{1}{\overline{\log} K} \min_{2\le k\le K} \left\{ \frac{1}{k} f\left(\nu_{\sigma_k}, \nu^\star\right) \right\}. \qquad \square$$

5.3.2. On Links between Φ and the Quantities $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\le}_{\inf}$, $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\ge}_{\inf}$

The Fenchel-Legendre transform $\phi^\star_\nu$ of the logarithmic moment-generating function of ν admits a classical (see, e.g., Boucheron et al., 2013, Exercice 4.13) dual formulation in terms of infima of Kullback-Leibler divergences. The following lemma, proved in Section 5.5.3 (see page 149), reveals that these infima correspond to $\mathcal{L}^{\le}_{\inf}$ and $\mathcal{L}^{\ge}_{\inf}$ for the model $\mathcal{P}[0,1]$ of distributions supported on $[0,1]$.

Lemma 5.5. *Consider the model $\mathcal{D} = \mathcal{P}[0,1]$. For all $\nu \in \mathcal{P}[0,1]$,*

$$\forall x \le \mathrm{E}(\nu), \quad \phi^\star_\nu(x) = \mathcal{L}^{\le}_{\inf}(x,\nu) \qquad \textit{and} \qquad \forall x \ge \mathrm{E}(\nu), \quad \phi^\star_\nu(x) = \mathcal{L}^{\ge}_{\inf}(x,\nu)\,.$$

Based on this lemma, we have the following rewriting, which is useful to reinterpret the quantities appearing in Theorem 5.3 and Corollary 5.4: $\Phi(\nu',\nu) = \mathcal{L}(\nu',\nu)$ for the model $\mathcal{P}[0,1]$, i.e.,

$$\inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]} \left\{\phi^\star_{\nu'}(x) + \phi^\star_\nu(x)\right\} = \inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]} \left\{\mathcal{L}^{\ge}_{\inf}(x,\nu') + \mathcal{L}^{\le}_{\inf}(x,\nu)\right\}. \tag{5.22}$$

For canonical one-parameter exponential models $\mathcal{D}_{\exp}$, a slightly weaker version of Lemma 5.5, only holding for x corresponding to expectations in $\mathcal{D}_{\exp}$ and provided in page 151 of 5.5.3.151, similarly shows (5.22), i.e., $\Phi = \mathcal{L}$. Conditions on general models for $\Phi = \mathcal{L}$ to hold are discussed in 5.5.3.152.

5.4. Lower Bounds

In most of this section, we restrict our attention to *generic*[2] K–armed bandit problems $\underline{\nu}$, that are such that $\mu_j \neq \mu_k$ for $j \neq k$. In particular, the best arm $a^\star(\underline{\nu})$ is unique.

Definition of a strategy, and of a (doubly-indexed) sequence of strategies. In the fixed-budget setting, a strategy is defined by

- a *sampling rule*, which consists in choosing the arm $A_t \in [K]$ to observe at each time step $t \in [T]$. This arm A_t depends on the previous observations $Y_1,\dots,Y_{t-1}$, but also possibly on some external randomization that we capture by the random variable U_{t-1}. A_t is thus $\mathcal{F}_{t-1}$–measurable, where

$$\mathcal{F}_{t-1} \overset{\text{def}}{=} \sigma(I_{t-1})\,, \qquad \text{with } I_{t-1} \overset{\text{def}}{=} \left(U_0, Y_1, U_1, Y_2, U_2, \dots, Y_{t-1}, U_{t-1}\right).$$

  I_{t-1} corresponds to the information available at the end of time step $t-1$, i.e., to the history.
- a *decision rule* $\hat{a}_T$ which is $\mathcal{F}_T$–measurable.

A strategy depends on the budget T and the number K of arms. As we are interested in asymptotic rates when the budget T goes to ∞, we consider sequences of strategies indexed by T. Our results will depend on assumptions made on the strategies, which will sometimes be stated for doubly-indexed sequences, that is, sequences of strategies indexed by T and K.

[2] This is probably a new terminology[3] for referring to bandit problems with no two same expectations for the distributions over the arms. It comes from measure theory: if expectations were drawn at random according to some diffuse distribution, e.g., a uniform distribution over an interval, or a Gaussian distribution, then, almost surely, no two expectations would be equal.

Outline of this section. As always in lower-bound results, there is a trade-off between how restrictive are the assumptions on the (doubly-indexed) sequences of strategies, and sometimes on the models, and how large the lower bounds are: the more restrictive the assumptions, the larger the lower bounds. We are interested in assumptions on strategies that are natural in the sense that they should be satisfied by `Successive-Rejects`-type strategies. For instance, Theorem 5.14 comes with the least assumptions but provides a bound where there are no divisions by the ranks k of the arms, which Theorems 5.10 and 5.13 do. We may see Theorem 5.10 as a warm-up result: its main aim is to generalize the lower bound by Audibert et al. (2010) to non-parametric models with a (non-constructive) proof that is only a few-line long. Our preferred result is Theorem 5.13, which provides the largest lower bound while putting the heaviest (though natural) constraints on the sequences of strategies.

5.4.1. Common Restriction: Consistence

For our lower bounds, we will consider sequences of strategies, either only indexed by $T \geq 1$ given a value of $K \geq 2$, or doubly indexed by T and K. These sequences will also be assumed to be "reasonable" in the sense below.

Consistent (or exponentially consistent) sequences of strategies. The misidentification probability $\mathbb{P}(\hat{a}_T \neq a^\star(\underline{\nu}))$ may vanish asymptotically (and even vanish exponentially fast) for all bandit problems —in not too large a model $\mathcal{D}$—, as illustrated in Section 5.3. We will therefore only be interested in such sequences of strategies, called (exponentially) consistent. In the sequel and for extra clarity, we index the probabilities by the ambient bandit problem $\underline{\nu}$ considered.

Definition 5.6. [(exponentially) consistant sequence of strategies]
Fix $K \geq 2$. A sequence of strategies indexed by $T \geq 1$ is consistent, respectively, exponentially consistent, on a model $\mathcal{D}$ if for all generic problems $\underline{\nu}$ in $\mathcal{D}$,

$$\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \underset{T\to+\infty}{\longrightarrow} 0\,, \qquad \text{respectively,} \qquad \limsup_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) < 0\,.$$

By extension, a doubly-indexed sequence of strategies is (exponentially) consistent if for all $K \geq 2$, the associated sequences of strategies are so.

The fundamental inequality. The fundamental inequality by Garivier et al. (2019), together with the very definition of consistency, yields in a straightforward manner our building block for lower bounds. Details of the derivation are provided in Section 2.3.3.

Lemma 5.7. *Fix $K \geq 2$ and a model $\mathcal{D}$. Consider a consistent sequence of strategies on $\mathcal{D}$, and two generic bandit problems $\underline{\nu}$ and $\underline{\lambda}$ in $\mathcal{D}$ such that $a^\star(\underline{\lambda}) \neq a^\star(\underline{\nu})$. Then*

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\limsup_{T\to+\infty} \sum_{a=1}^{K} \frac{\mathbb{E}_{\underline{\lambda}}[N_a(T)]}{T} \operatorname{KL}(\lambda_a, \nu_a)\,.$$

Note that the proof reveals that the inequality actually holds for limits taken along sub-sequences $(T_n)_{n\geq 1}$. Also, we may only relax the assumptions on the bandit models; e.g., they do not need to be generic and it suffices that they have different unique optimal arms (the notion of a generic bandit problem is defined in the first lines of Section 5.4).

5.4.2. A Lower Bound Revisiting and Extending the one by Audibert et al. (2010)

The focus of this subsection is to establish the lower bound (5.5), from which we derived the gap-based lower bound (5.6) by Audibert et al. (2010). The lower bound (5.5) is smaller than the lower bound to be exhibited in the next subsection, but it comes with less restrictive assumptions on the behaviors of the sequences of strategies considered.

Firstly, we only consider sequences of strategies —actually, sequences of sampling schemes— that do not pull too often the worst arm, and which we will refer to as being balanced against the worst arm. `Successive-Rejects`-type strategies sample the worst arm less than other arms in expectations, and hence, are indeed balanced against the worst arm. To define this constraint formally, we denote by $w_\star(\underline{\nu})$ the index of the unique worst arm of a generic bandit problem $\underline{\nu}$.

Definition 5.8. [balanced against the worst arm sequence]
A doubly-indexed sequence of strategies is *balanced against the worst arm* on a model $\mathcal{D}$ if for all $K \geq 2$, for all generic K–armed bandit problems $\underline{\nu}$ in $\mathcal{D}$,

$$\limsup_{T\to+\infty} \frac{1}{T}\,\mathbb{E}_{\underline{\nu}}[N_{w_\star(\underline{\nu})}(T)] \leq \frac{1}{K}\,.$$

A second constraint is related to bandit sub-problems. We say that $\underline{\nu}'$ is a sub-problem of a K–armed bandit problem $\underline{\nu}$ if $\underline{\nu}' = (\nu_a)_{a\in\mathcal{A}}$ for a subset $\mathcal{A} \subseteq [K]$ of cardinality greater than or equal to 2; we denote by $\underline{\nu}' \subseteq \underline{\nu}$ this fact. We say in addition that $\underline{\nu}'$ and $\underline{\nu}$ feature the same optimal arm if $\nu'_{a^\star(\underline{\nu}')} = \nu_{a^\star(\underline{\nu})}$. It should be easier to identify the best arm in $\underline{\nu}'$ than in $\underline{\nu}$, in the sense below, and this defines the fact that a strategy cleverly exploits pruning of sub-optimal arms. Again, `Successive-Rejects`-type strategies naturally satisfy this constraint.

Definition 5.9. [clever exploitation of the pruning of sub-optimal arms]
A doubly-indexed sequence of strategies *cleverly exploits pruning of sub-optimal arms* on a model $\mathcal{D}$ if for all generic bandit problems $\underline{\nu}$ in $\mathcal{D}$ with $K \geq 2$ arms, for all sub-problems $\underline{\nu}' \subseteq \underline{\nu}$ featuring the same optimal arm,

$$\liminf_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq \liminf_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}'}(\hat{a}_T \neq a^\star(\underline{\nu}'))\,.$$

We use again the order statistics $\mu_{w_\star(\underline{\nu})} = \mu_{(K)} < \mu_{(K-1)} < \ldots < \mu_{(1)} = \mu_{a^\star(\underline{\nu})}$.

Theorem 5.10. *Fix a model $\mathcal{D}$. Consider a doubly-indexed sequence of strategies that is consistent, balanced against the worst arm on $\mathcal{D}$, and that cleverly exploits the pruning of sub-optimal arms on $\mathcal{D}$. For all generic bandit problems $\underline{\nu}$ in $\mathcal{D}$ with $K \geq 2$ arms,*

$$\liminf_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\min_{2\leq k\leq K} \frac{\mathcal{L}^<_{\inf}(\mu_{(k)},\nu^\star)}{k}\,.$$

Proof. The proof consists of two steps. The first step is to prove that for a generic bandit problem $\underline{\nu}$ in $\mathcal{D}$ with $K \geq 2$ arms, we have,

$$\liminf_{T\to+\infty} \frac{1}{T}\log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\frac{\mathcal{L}^<_{\inf}(\mu_{(K)},\nu^\star)}{K}\,. \tag{5.23}$$

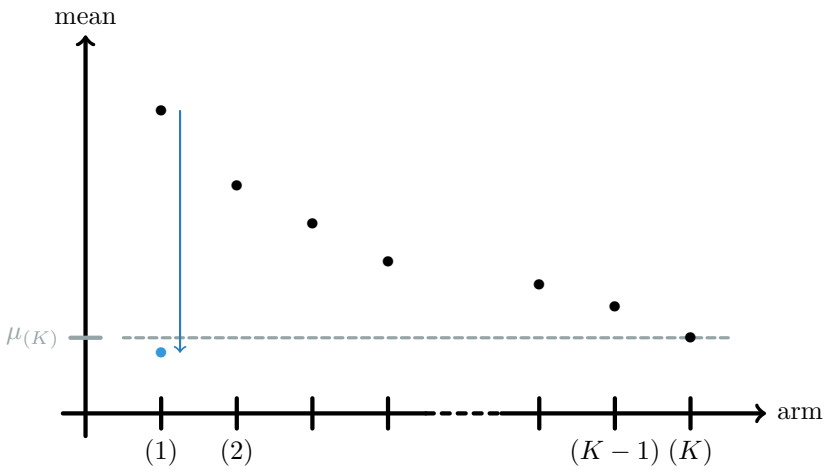


Figure 5.4: The alternative $\underline{\lambda}$ of bandit problem $\underline{\nu}$ is obtained by modifying only the best arm of $\underline{\nu}$.

In the second step, we use this lower bound and the very definition of the clever exploitation of the pruning of sub-optimal arms to get the claimed bound.

Step 1: lower bound (5.23)**.** We follow a well-established methodology and consider an alternative bandit problem only differing from $\underline{\nu}$ at one arm, namely, at the best arm. To do so, we set some distribution $\zeta \in \mathcal{D}$ with $\mathrm{E}(\zeta) < \mu_{(K)}$, if some exists, and define the bandit problem $\underline{\lambda} = (\lambda_1, \dots, \lambda_K)$ as

$$\lambda_a = \begin{cases} \zeta & \text{if } a = a^\star(\underline{\nu}), \\ \nu_a & \text{if } a \neq a^\star(\underline{\nu}). \end{cases}$$

Observe (see Figure 5.4 that $\underline{\lambda}$ is also a generic bandit problem in $\mathcal{D}$, that $a^\star(\underline{\nu})$ is the worst arm in $\underline{\lambda}$ (and also that the second best arm of $\underline{\nu}$ is the optimal arm in $\underline{\lambda}$, but we will not use this specific fact). Therefore, Lemma 5.7 yields, as $\underline{\lambda}$ and $\underline{\nu}$ only differ at arm $a^\star(\underline{\nu})$,

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\limsup_{T\to+\infty} \frac{\mathbb{E}_{\underline{\lambda}}[N_{a^\star(\underline{\nu})}(T)]}{T} \, \mathrm{KL}(\lambda_{a^\star(\underline{\nu})}, \nu^\star) \,,$$

where we recall that $\nu^\star = \nu_{a^\star(\underline{\nu})}$. Given that $a^\star(\underline{\nu})$ is the worst arm of $\underline{\lambda}$, and since by assumption, the sequence of strategies is balanced against the worst arm,

$$\limsup_{T\to+\infty} \frac{1}{T} \, \mathbb{E}_{\underline{\lambda}}[N_{a^\star(\underline{\nu})}(T)] \leq \frac{1}{K} \,,$$

proving that

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\frac{\mathrm{KL}(\zeta, \nu^\star)}{K} \,.$$

The claimed inequality (5.23) follows from taking the supremum on the right-hand side over distributions $\zeta \in \mathcal{D}$ with $\mathrm{E}(\zeta) < \mu_{(K)}$.

Step 2: clever exploitation of pruning. For each $k \in \{2, \dots, K-1\}$, define $\underline{\nu}'_{1:k}$ as the sub-problem of $\underline{\nu}$ obtained by keeping the k best arms and dropping the $K-k$ worst arms. Use the definition of clever exploitation of pruning of sub-optimal arms and apply (5.23) to $\underline{\nu}'_{1:k}$ to get

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq \liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}'_{1:k}}\Big(\hat{a}_T \neq a^\star(\underline{\nu}'_{1:k})\Big) \geq -\frac{\mathcal{L}^<_{\inf}(\mu_{(k)}, \nu^\star)}{k} \,.$$

Taking the maximum of all lower bounds exhibited as k varies between 2 and K, we proved the claimed result. □

5.4.3. A Larger Lower Bound, for a more Restrictive Class of Strategies

In this section, we derive a slightly stronger version of the lower bound (5.4). This lower bound is larger than the bound exhibited in the previous subsection but relies on stronger assumptions on the strategies considered. Namely, we introduce an assumption of monotonicity, which extends Definition 5.8 to provide frequency constraints on each arm $a \in [K]$.

Definition 5.11. **[monotonous sequence of strategies]**
Fix $K \geq 2$. A sequence of strategies is *monotonous* on a model $\mathcal{D}$ if for all generic problems $\underline{\nu}$ in $\mathcal{D}$, for all arms $a \in \{1, \dots, K\}$,

$$\limsup_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_{(a)}(T)]}{T} \leq \frac{1}{a},$$

where arms are ordered such that $\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$.

This condition is satisfied as soon as a given arm is not pulled more often, asymptotically and on average, than better-performing arms (note that Definition 5.11 is slightly weaker than this). `Successive-Rejects`-type strategies naturally satisfy this requirement.

We also rely on the following assumption on the model $\mathcal{D}$, which essentially indicates that there is "no gap" in $\mathcal{D}$. Once again, the model $\mathcal{P}[0,1]$ and canonical one-parameter exponential models $\mathcal{D}_{\exp}$ all satisfy this mild requirement (see Section 5.5.4 for the immediate details).

Definition 5.12. A model $\mathcal{D}$ is *normal* if for all $\nu \in \mathcal{D}$, for all $x \geq \mathrm{E}(\nu)$,

$$\begin{aligned} \forall \varepsilon > 0, \qquad \mathcal{L}^{>}_{\inf}(x, \nu) &\stackrel{\text{def}}{=} \inf\{\mathrm{KL}(\zeta, \nu) : \ \zeta \in \mathcal{D} \ \text{ s.t. } \ \mathrm{E}(\zeta) > x\} \\ &= \inf\{\mathrm{KL}(\zeta, \nu) : \ \zeta \in \mathcal{D} \ \text{ s.t. } \ x + \varepsilon > \mathrm{E}(\zeta) > x\}. \end{aligned}$$

Theorem 5.13. *Fix $K \geq 2$ and a normal model $\mathcal{D}$. Consider a sequence of strategies that is consistent and monotonous on $\mathcal{D}$. For all generic bandit problems $\underline{\nu}$ in $\mathcal{D}$,*

$$\liminf_{T \to +\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{2 \leq k \leq K} \min_{2 \leq j \leq k} \inf_{x \in [\mu_{(j)}, \mu_{(j-1)})} \left\{ \frac{\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})}{j-1} + \frac{\mathcal{L}^{<}_{\inf}(x, \nu^\star)}{j} \right\}.$$

Proof. We fix a generic bandit $\underline{\nu}$ in $\mathcal{D}$ and consider the following sets of alternative bandit problems, indexed by triplets (k, j, x) satisfying $2 \leq k \leq K$ and $2 \leq j \leq k$, as well as $x \in [\mu_{(j)}, \mu_{(j-1)})$:

$$\mathrm{Alt}_{k,j,x}(\underline{\nu}) = \left\{ \underline{\lambda} \text{ in } \mathcal{D} : \ \mathrm{E}(\lambda_{(1)}) < x < \mathrm{E}(\lambda_{(k)}) < \mu_{(j-1)} \ \text{ and } \ \lambda_a = \nu_a \ \text{ for } \ a \notin \{(1), (k)\} \right\};$$

in particular, an alternative problem $\underline{\lambda}$ in $\mathrm{Alt}_{k,j,x}(\underline{\nu})$ only differ from the original bandit problem $\underline{\nu}$ at the best arm (1) and at the k–th best arm (k). Given $x \in [\mu_{(j)}, \mu_{(j-1)})$ and $\mathrm{E}(\lambda_{(1)}) < x$, arm (1) is at best the j–th best arm of $\underline{\lambda}$, but it can be possibly worse. Similarly, the same condition on x and the fact that $x < \mathrm{E}(\lambda_{(k)})$ implies that arm (k) is exactly the $j-1$–th best arm of $\underline{\lambda}$. Both facts are illustrated in Figure 5.5.

Thus, by monotonicity of the strategy,

$$\limsup_{T \to +\infty} \frac{\mathbb{E}_{\underline{\lambda}}[N_{(k)}(T)]}{T} \leq \frac{1}{j-1} \qquad \text{and} \qquad \limsup_{T \to +\infty} \frac{\mathbb{E}_{\underline{\lambda}}[N_{(1)}(T)]}{T} \leq \frac{1}{j}.$$

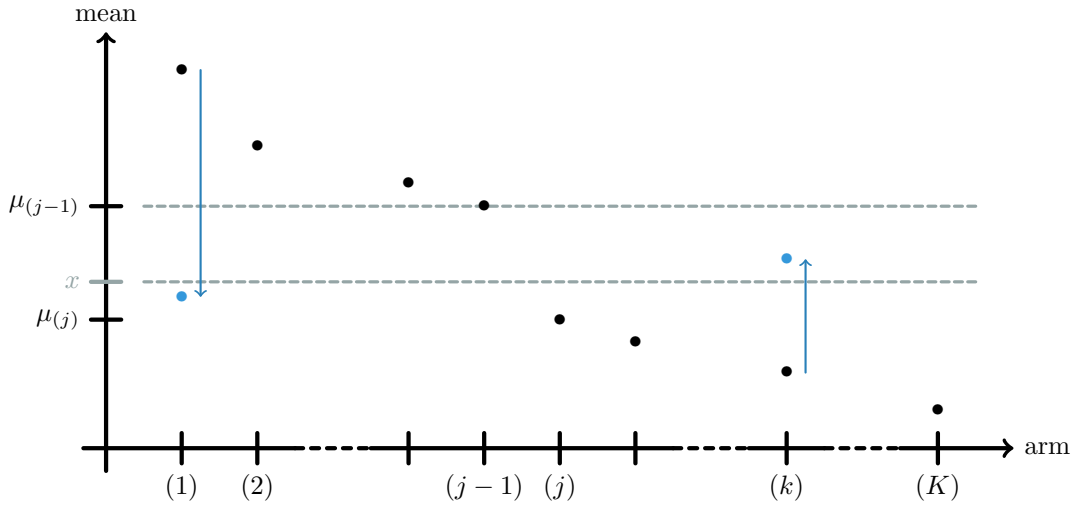


Figure 5.5: Original bandit problem $\underline{\nu}$ (in dark) and modifications made to arms (1) and (k) to obtain an alternative bandit problem $\underline{\lambda} \in \mathrm{Alt}_{k,j,x}(\underline{\nu})$ (in blue): in $\underline{\lambda}$, arm (k) is the $j-1$–th best arm, while arm $(1) = a^\star(\underline{\nu})$ is at best the j–th best arm.

Given that the optimal arm in $\underline{\lambda}$ is different from the optimal arm (1) of $\underline{\nu}$, Lemma 5.7 may be applied; together with the two upper bounds above, it yields

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\left(\frac{\mathrm{KL}(\lambda_{(k)}, \nu_{(k)})}{j-1} + \frac{\mathrm{KL}(\lambda_{(1)}, \nu^\star)}{j} \right).$$

We can now take the infimum over all bandit problems $\underline{\lambda} \in \mathrm{Alt}_{k,j,x}(\underline{\nu})$ and obtain the following lower bound, where we define a quantity $\mathcal{I}_{k,j,x}(\underline{\nu})$:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\inf_{\underline{\lambda}\in\mathrm{Alt}_{k,j,x}(\underline{\nu})} \left\{ \frac{\mathrm{KL}(\lambda_{(k)}, \nu_{(k)})}{j-1} + \frac{\mathrm{KL}(\lambda_{(1)}, \nu^\star)}{j} \right\} \stackrel{\mathrm{def}}{=} -\mathcal{I}_{k,j,x}(\underline{\nu}).$$

We prove below that

$$\mathcal{I}_{k,j,x}(\underline{\nu}) = \frac{\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})}{j-1} + \frac{\mathcal{L}^{<}_{\inf}(x, \nu^\star)}{j}, \tag{5.24}$$

from which the lower bound claimed in Theorem 5.13 will follow, by taking the supremum of $-\mathcal{I}_{k,j,x}(\underline{\nu})$ first over $x \in [\mu_{(j)}, \mu_{(j-1)})$, then the maximum over $2 \leq j \leq k$, and finally, the maximum over $2 \leq k \leq K$.

We now prove (5.24). The infimum over $\underline{\lambda} \in \mathrm{Alt}_{k,j,x}(\underline{\nu})$ may be split into two separate infima, respectively over $\lambda_{(k)}$ and $\lambda_{(1)}$; given that each term of the sum of KL only depends either on $\lambda_{(k)}$, or on $\lambda_{(1)}$, but not on both, we may write

$$\begin{aligned}
\mathcal{I}_{k,j,x}(\underline{\nu}) &= \inf_{\substack{\lambda_{(1)},\lambda_{(k)}\in\mathcal{D}:\\ \mathrm{E}(\lambda_{(1)})<x\\ x<\mathrm{E}(\lambda_{(k)})<\mu_{(j-1)}}} \left\{ \frac{\mathrm{KL}(\lambda_{(k)}, \nu_{(k)})}{j-1} + \frac{\mathrm{KL}(\lambda_{(1)}, \nu^\star)}{j} \right\} \\
&= \frac{1}{j-1} \underbrace{\inf_{\substack{\lambda_{(k)}\in\mathcal{D}:\\ x<\mathrm{E}(\lambda_{(k)})<\mu_{(j-1)}}} \mathrm{KL}(\lambda_{(k)}, \nu_{(k)})}_{=\mathcal{L}^{>}_{\inf}(x,\nu_{(k)})} + \frac{1}{j} \underbrace{\inf_{\substack{\lambda_{(1)}\in\mathcal{D}:\\ \mathrm{E}(\lambda_{(1)})<x}} \mathrm{KL}(\lambda_{(1)}, \nu^\star)}_{=\mathcal{L}^{<}_{\inf}(x,\nu^\star)},
\end{aligned}$$

where we obtained $\mathcal{L}^{<}_{\inf}(x, \nu^\star)$ by definition while we relied on the normality of the model (Definition 5.12) to obtain $\mathcal{L}^{>}_{\inf}(x, \nu_{(k)})$. We did so with $\varepsilon = \mu_{(j-1)} - x$, which is indeed positive as we considered $x < \mu_{(j-1)}$. □

5.4.4. A General Lower Bound, Valid for any Strategy

The previous subsections illustrated what may be achieved under restrictions —though natural restrictions— on the classes of strategies considered. For the sake of completeness, we also provide a lower bound relying on no other restriction than consistency; it extends the lower bound (5.9) exhibited by Kaufmann et al. (2016) for $K = 2$ arms, and is formulated in terms of $\mathcal{L}^{<}_{\inf}$ and $\mathcal{L}^{>}_{\inf}$.

Theorem 5.14. *Fix $K \geq 2$ and a model $\mathcal{D}$. Consider a consistent sequence of strategies on $\mathcal{D}$. For all generic bandit problems $\underline{\nu}$ in $\mathcal{D}$,*

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \min_{k\neq a^\star(\underline{\nu})} \inf_{x\in[\mu_k,\mu^\star]} \max\{\mathcal{L}^{>}_{\inf}(x,\nu_k), \mathcal{L}^{<}_{\inf}(x,\nu^\star)\}\,.$$

Proof. Let $\underline{\nu}$ be a generic bandit problem. We fix $k \neq a^\star(\underline{\nu})$ and $x \in [\mu_k, \mu^\star]$, and prove that

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \max\{\mathcal{L}^{>}_{\inf}(x,\nu_k), \mathcal{L}^{<}_{\inf}(x,\nu^\star)\}\,,$$

from which the stated lower bound follows, by taking suprema. To do so, we consider the set of alternative bandit problems

$$\mathrm{Alt}_{k,x}(\underline{\nu}) = \Big\{\underline{\lambda} \text{ in } \mathcal{D}:\ \mathrm{E}(\lambda_{a^\star(\underline{\nu})}) < x < \mathrm{E}(\lambda_k) \ \text{ and } \ \lambda_a = \nu_a \ \text{ for } \ a \notin \{a^\star(\underline{\nu}), k\}\Big\}\,;$$

it is composed of bandit problems, only differing from $\underline{\nu}$ at arms $a^\star(\underline{\nu})$ and k, and for which arm k is better than arm $a^\star(\underline{\nu})$, with associated expectations separated by x. In particular, the optimal arm in $\underline{\lambda}$ is different from the optimal arm $a^\star(\underline{\nu})$ of $\underline{\nu}$. Lemma 5.7 may therefore be applied; it states that

$$\begin{aligned}
&\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \\
&\geq - \limsup_{T\to+\infty} \frac{\mathbb{E}_{\underline{\lambda}}[N_k(T)]}{T}\ \mathrm{KL}(\lambda_k,\nu_k) + \frac{\mathbb{E}_{\underline{\lambda}}[N_{a^\star(\underline{\nu})}(T)]}{T}\ \mathrm{KL}(\lambda_{a^\star(\underline{\nu})},\nu_{a^\star(\underline{\nu})}) \\
&\geq - \max\Big\{\mathrm{KL}(\lambda_k,\nu_k),\ \mathrm{KL}(\lambda_{a^\star(\underline{\nu})},\nu_{a^\star(\underline{\nu})})\Big\}\,,
\end{aligned}$$

where we used, for the second inequality, the crude upper bound $N_k(T) + N_{a^\star(\underline{\nu})}(T) \leq T$. Taking the supremum of the obtained lower bound over all $\underline{\lambda} \in \mathrm{Alt}_{k,x}(\underline{\nu})$ leads to the following inequality, where we define the short-hand notation $\mathcal{I}_{k,x}(\underline{\nu})$:

$$\liminf_{T\to+\infty} \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq - \inf_{\underline{\lambda}\in\mathrm{Alt}_{k,x}(\underline{\nu})} \max\Big\{\mathrm{KL}(\lambda_k,\nu_k), \mathrm{KL}(\lambda_{a^\star(\underline{\nu})},\nu_{a^\star(\underline{\nu})})\Big\} \stackrel{\text{def}}{=} -\mathcal{I}_{k,x}(\underline{\nu})\,.$$

The proof is concluded below by showing that $\mathcal{I}_{k,x}(\underline{\nu}) = \max\{\mathcal{L}^{>}_{\inf}(x,\nu_k), \mathcal{L}^{<}_{\inf}(x,\nu^\star)\}$.

As in the proof of Theorem 5.13, we use a separation of the infima, in the abstract form, for two functions f and g,

$$\inf_{u,v}\ \max\{f(u),\, g(v)\} = \max\Big\{\inf_u f(u)\,,\ \inf_v g(v)\Big\}\,.$$

Here, by definition of $\mathrm{Alt}_{k,x}(\underline{\nu})$,

$$\begin{aligned}\mathcal{I}_{k,x}(\underline{\nu}) &= \inf_{\substack{\lambda_{a^\star(\underline{\nu})},\lambda_k\in\mathcal{D}\\ \mathrm{E}(\lambda_{a^\star(\underline{\nu})})<x\\ \mathrm{E}(\lambda_k)>x}} \max\Big\{\mathrm{KL}(\lambda_k,\nu_k),\ \mathrm{KL}(\lambda_{a^\star(\underline{\nu})},\nu_{a^\star(\underline{\nu})})\Big\}\\ &= \max\left\{\inf_{\substack{\lambda_k\in\mathcal{D}\\ \mathrm{E}(\lambda_k)>x}} \mathrm{KL}(\lambda_k,\nu_k),\quad \inf_{\substack{\lambda_{a^\star(\underline{\nu})}\in\mathcal{D}\\ \mathrm{E}(\lambda_{a^\star(\underline{\nu})})<x}} \mathrm{KL}(\lambda_{a^\star(\underline{\nu})},\nu_{a^\star(\underline{\nu})})\right\}\\ &= \max\Big\{\mathcal{L}^{>}_{\inf}(x,\nu_k),\mathcal{L}^{<}_{\inf}(x,\nu^\star)\Big\},\end{aligned}$$

which concludes the proof. □

5.5. Technical Details

5.5.1. Properties of the $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$

General Properties

We present here general properties, that hold for all models $\mathcal{D}$, of the $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$, and $\mathcal{L}^{\geq}_{\inf}$ quantities. We state some properties for $\mathcal{L}^{<}_{\inf}$, that all also hold for $\mathcal{L}^{\leq}_{\inf}$; the corresponding properties for $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$ are deduced by symmetry.

The function $\mathcal{L}^{<}_{\inf}(\,\cdot\,,\nu)$ is non-increasing and satisfies $\mathcal{L}^{<}_{\inf}(x,\nu)=0$ for all $x>\mathrm{E}(\nu)$, as can be seen by taking $\zeta=\nu$. Also, whenever $\mathcal{D}$ is convex, the function $\mathcal{L}^{<}_{\inf}$ is jointly convex over $\mathbb{R}\times\mathcal{D}$, as indicated in the lemma below. In particular, $x\mapsto\mathcal{L}^{<}_{\inf}(x,\nu)$ is continuous on the interior of its domain (the set where it takes finite values).

Lemma 5.15. *When $\mathcal{D}$ is a convex model, all four functions $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$, and $\mathcal{L}^{\geq}_{\inf}$ are jointly convex over $\mathbb{R}\times\mathcal{D}$.*

Proof. We provide the proof for $\mathcal{L}^{<}_{\inf}$, and it may be adapted in a straightforward manner for the other functions.

We set two distributions ν and ν' of $\mathcal{D}$, two expectation levels μ and μ' in $\mathbb{R}$, and a weight $\lambda\in(0,1)$. We want to prove that

$$\mathcal{L}^{<}_{\inf}(\lambda\mu+(1-\lambda)\mu',\lambda\nu+(1-\lambda)\nu')\leq\lambda\mathcal{L}^{<}_{\inf}(\mu,\nu)+(1-\lambda)\mathcal{L}^{<}_{\inf}(\mu',\nu')\,. \tag{5.25}$$

The desired inequality holds whenever $\mathcal{L}^{<}_{\inf}(\mu,\nu)=+\infty$ or $\mathcal{L}^{<}_{\inf}(\mu',\nu')=+\infty$. Otherwise, assuming that both $\mathcal{L}^{<}_{\inf}(\mu,\nu)$ and $\mathcal{L}^{<}_{\inf}(\mu',\nu')$ are finite, we set $\delta>0$ (which we will ultimately let converge to 0) and pick ζ and ζ' in $\mathcal{D}$ such that $\mathrm{E}(\zeta)<\mu$ and $\mathrm{E}(\zeta)<\mu'$, as well as

$$\mathrm{KL}(\zeta,\nu)\leq\mathcal{L}^{<}_{\inf}(\mu,\nu)+\delta\quad\text{and}\quad\mathrm{KL}(\zeta',\nu')\leq\mathcal{L}^{<}_{\inf}(\mu',\nu')+\delta\,.$$

Then, by joint convexity of the Kullback-Leibler divergence:

$$\begin{aligned}\lambda\mathcal{L}^{<}_{\inf}(\mu,\nu)+(1-\lambda)\mathcal{L}^{<}_{\inf}(\mu',\nu')+\delta &\geq \lambda\,\mathrm{KL}(\zeta,\nu)+(1-\lambda)\,\mathrm{KL}(\zeta',\nu')\\ &\geq \mathrm{KL}(\lambda\zeta+(1-\lambda)\zeta',\lambda\nu+(1-\lambda)\nu')\\ &\geq \mathcal{L}^{<}_{\inf}(\lambda\mu+(1-\lambda)\mu',\lambda\nu+(1-\lambda)\nu')\,,\end{aligned}$$

where for the last inequality, we used the definition of $\mathcal{L}^{<}_{\inf}$ as an infimum and the fact that by convexity, the distribution $\lambda\zeta+(1-\lambda)\zeta'$ belongs to $\mathcal{D}$, with expectation larger than $\lambda\mu+(1-\lambda)\mu'$. The desired convexity inequality (5.25) follows by letting $\delta\to0$. □

Specific Properties for the Model $\mathcal{P}[0,1]$

We study some properties of the $\mathcal{L}_{\inf}^{<}, \mathcal{L}_{\inf}^{\leq}, \mathcal{L}_{\inf}^{>}$ and $\mathcal{L}_{\inf}^{\geq}$ quantities for the model $\mathcal{P}[0,1]$ of all distributions over $[0,1]$. Similar analysis might be done for, e.g., exponential models, as explained in Section 5.5.3.

Since we are considering distributions over the interval $[0,1]$, the data-processing inequality (2.11) for Kullback-Leibler divergences ensures that for all $\zeta \in \mathcal{P}[0,1]$,

$$\mathrm{KL}(\zeta,\nu) \geq \mathrm{KL}\Big(\mathrm{Ber}(\mathrm{E}(\zeta)),\, \mathrm{Ber}(\mathrm{E}(\nu))\Big) \geq 2\big(\mathrm{E}(\zeta)-\mathrm{E}(\nu)\big)^2\,,$$

where $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p and where we applied Pinsker's inequality for Bernoulli distributions. Therefore, taking the infimum over distributions $\zeta \in \mathcal{P}[0,1]$ with $\mathrm{E}(\zeta) < x$,

$$\forall x \leq \mathrm{E}(\nu), \quad \mathcal{L}_{\inf}^{<}(x,\nu) \geq 2\big(\mathrm{E}(\nu)-x\big)^2\,. \tag{5.26}$$

We denote by $m(\nu) = \min(\mathrm{Supp}(\nu)) \geq 0$ the minimum of the closed support $\mathrm{Supp}(\nu)$ of ν; that is, $m(\nu)$ is the largest value such that $\mathrm{Supp}(\nu) \subseteq \big[m(\nu),\,1\big]$. We will refer to $m(\nu)$ as the lower end of the support of ν. Though we will not need it immediately, we also define the upper end of the support of ν as $M(\nu) = \max(\mathrm{Supp}(\nu)) \leq 1$; by symmetry, it will be considered when studying $\mathcal{L}_{\inf}^{>}$ and $\mathcal{L}_{\inf}^{\geq}$ instead of $\mathcal{L}_{\inf}^{<}$ and $\mathcal{L}_{\inf}^{\leq}$.

The lemma below states that the functions $\mathcal{L}_{\inf}^{<}(\,\cdot\,,\nu)$ and $\mathcal{L}_{\inf}^{\leq}(\,\cdot\,,\nu)$ coincide, except maybe at $m(\nu)$. One may wonder what happens at $x = m(\nu)$. We denote by $\nu\big\{m(\nu)\big\}$ the probability mass assigned by ν to the point $m(\nu)$. It follows from the second part of the lemma below that $\mathcal{L}_{\inf}^{<}(m(\nu),\nu) = \mathcal{L}_{\inf}^{\leq}(m(\nu),\nu)$ if and only if $\big\{m(\nu)\big\}$ is not an atom of ν.

Lemma 5.16. *We consider the model $\mathcal{D} = \mathcal{P}[0,1]$. The function $\mathcal{L}_{\inf}^{<}(\,\cdot\,,\nu)$ is continuous on the interval $\big(m(\nu),\,+\infty\big)$. We also have, on the one hand,*

$$\forall \mu \neq m(\nu), \qquad \mathcal{L}_{\inf}^{<}(\mu,\nu) = \mathcal{L}_{\inf}^{\leq}(\mu,\nu)\,, \tag{5.27}$$

and on the other hand, at $\mu = m(\nu)$,

$$\log\frac{1}{\nu\big\{m(\nu)\big\}} = \mathcal{L}_{\inf}^{\leq}(m(\nu),\nu) \leq \mathcal{L}_{\inf}^{<}(m(\nu),\nu) = +\infty\,. \tag{5.28}$$

Analogous results hold for $\mathcal{L}_{\inf}^{>}(\,\cdot\,,\nu)$, $\mathcal{L}_{\inf}^{\geq}(\,\cdot\,,\nu)$, and $M(\nu)$.

Proof. To prove (5.27), we first identify the interior of the domain of $\mathcal{L}_{\inf}^{<}$.

Distributions ζ such that $\mathrm{E}(\zeta) < m(\nu)$ cannot be absolutely continuous with respect to ν; otherwise, they would also give a null probability to values strictly smaller than $m(\nu)$, which contradicts the assumption $\mathrm{E}(\zeta) < m(\nu)$. Hence $\mathrm{KL}(\zeta,\nu) = +\infty$ for these distributions. It follows that $\mathcal{L}_{\inf}^{<}(\mu,\nu) = \mathcal{L}_{\inf}^{\leq}(\mu,\nu) = +\infty$ for $\mu < m(\nu)$; we note in passing that we also have $\mathcal{L}_{\inf}^{<}(m(\nu),\nu) = +\infty$.

For $\mu > m(\nu)$, we take $\varepsilon > 0$ with $m(\nu)+\varepsilon < \mu$ and have, by definition of the support of a measure, that $\big[m(\nu), m(\nu)+\varepsilon\big]$ has a positive ν–measure denoted by κ. The distribution ζ given by ν conditioned to the interval $\big[m(\nu), m(\nu)+\varepsilon\big]$ is absolutely continuous with respect to ν, with density $\mathrm{d}\zeta/\mathrm{d}\nu = 1/\kappa$ on $\big[m(\nu), m(\nu)+\varepsilon\big]$, and 0 elsewhere; therefore, $\mathrm{KL}(\zeta,\nu) = \log(1/\kappa) < +\infty$ and $\mathcal{L}_{\inf}^{<}(\mu,\nu) < +\infty$.

The interior of the domain of $\mu \mapsto \mathcal{L}_{\inf}^{<}(\mu,\nu)$ is therefore $\big(m(\nu),\,+\infty\big)$, and we recall that $\mathcal{L}_{\inf}^{<}(\,\cdot\,,\nu)$ is continuous on this interval. We fix some $\mu > m(\nu)$. For all $\varepsilon > 0$, by the very

definitions of all quantities as infima of nested sets, we have

$$\mathcal{L}^{<}_{\inf}(\mu-\varepsilon,\nu) \le \mathcal{L}^{\le}_{\inf}(\mu,\nu) \le \mathcal{L}^{<}_{\inf}(\mu,\nu)\,.$$

Letting $\varepsilon \to 0$, we get, by a sandwich argument, that $\mathcal{L}^{\le}_{\inf}(\mu,\nu) = \mathcal{L}^{<}_{\inf}(\mu,\nu)$. This concludes the proof of (5.27).

We turn our attention to (5.28). We already showed above that $\mathcal{L}^{<}_{\inf}(m(\nu),\nu) = +\infty$. Now, to compute $\mathcal{L}^{\le}_{\inf}(\mu,\nu)$, we wonder which are the distributions ζ that are absolutely continuous with respect to ν, and thus, give a null probability to values strictly smaller than $m(\nu)$, and are also such that $\mathrm{E}(\zeta) \le m(\nu)$: at most one such distribution exists, the Dirac mass at $m(\nu)$, denoted by $\delta_{m(\nu)}$. We then distinguish the cases $\nu\{m(\nu)\} > 0$ and $\nu\{m(\nu)\} = 0$ to establish, respectively, the equalities

$$\mathcal{L}^{\le}_{\inf}(m(\nu),\nu) = \mathrm{KL}(\delta_{m(\nu)},\nu) = \log\frac{1}{\nu\{m(\nu)\}} \quad \text{and} \quad \mathcal{L}^{\le}_{\inf}(m(\nu),\nu) = +\infty = \log\frac{1}{\nu\{m(\nu)\}}\,.$$

In both cases, the first equality in (5.28) is proved, which concludes the proof. □

We also have the following result, which is the most important and useful one, as it discussed the quantity that appears in the upper bounds on the average log-probability of misidentification of the optimal arm; see Corollary 5.4 together with Lemma 5.5.

Lemma 5.17. *Let $\nu,\nu' \in \mathcal{P}[0,1]$ with $\mu = \mathrm{E}(\nu) > \mathrm{E}(\nu') = \mu'$. Then*

$$\inf_{x\in[\mu',\mu]} \mathcal{L}^{\le}_{\inf}(x,\nu) + \mathcal{L}^{\ge}_{\inf}(x,\nu') = \inf_{x\in[\mu',\mu]} \mathcal{L}^{<}_{\inf}(x,\nu) + \mathcal{L}^{>}_{\inf}(x,\nu')$$

if and only if either $m(\nu) \neq M(\nu')$ or $\nu\{m(\nu)\} \times \nu'\{M(\nu')\} = 0$.

Remark. In other words, the only case for which the two infima differ is when $m(\nu) = M(\nu')$, i.e., the upper end of the support of ν' equals the lower end of the support of ν, and both ν and ν' admit this common value as an atom.

Proof. The first lines of the proof of Lemma 5.16 show that $\mathcal{L}^{\le}_{\inf}(x,\nu) = \mathcal{L}^{<}_{\inf}(x,\nu) = +\infty$ for $x < m(\nu)$. We can symmetrically show that $\mathcal{L}^{\ge}_{\inf}(x,\nu') = \mathcal{L}^{>}_{\inf}(x,\nu') = +\infty$ for $x > M(\nu')$. Therefore, $\mathcal{L}^{\le}_{\inf}(x,\nu) + \mathcal{L}^{\ge}_{\inf}(x,\nu')$ and $\mathcal{L}^{<}_{\inf}(x,\nu) + \mathcal{L}^{>}_{\inf}(x,\nu')$ are infinite whenever x lies outside of $[m(\nu), M(\nu')]$. This implies that

$$\inf_{x\in[\mu',\mu]} \mathcal{L}^{\le}_{\inf}(x,\nu) + \mathcal{L}^{\ge}_{\inf}(x,\nu') = \inf_{x\in[\mu',\mu]\cap[m(\nu),M(\nu')]} \mathcal{L}^{\le}_{\inf}(x,\nu) + \mathcal{L}^{\ge}_{\inf}(x,\nu')$$

and

$$\inf_{x\in[\mu',\mu]} \mathcal{L}^{<}_{\inf}(x,\nu) + \mathcal{L}^{>}_{\inf}(x,\nu') = \inf_{x\in[\mu',\mu]\cap[m(\nu),M(\nu')]} \mathcal{L}^{<}_{\inf}(x,\nu) + \mathcal{L}^{>}_{\inf}(x,\nu')\,.$$

We now split the analysis according to how large the interval $\mathcal{I}$ is, where

$$\mathcal{I} = [\mu',\mu] \cap [m(\nu), M(\nu')] = \Big[\max\{\mu', m(\nu)\},\, \min\{\mu, M(\nu')\}\Big]\,.$$

Case 1: $\mathcal{I}$ is empty. In that case, the two infima are over an empty set and both equal $+\infty$.

Case 2: $\mathcal{I}$ has a non-empty interior. When $a \neq b$, the infimum of a convex function over a closed interval $[a,b]$ equals the infimum over (a,b), whether the function takes finite or infinite values at a and b. Now, the interior of $\mathcal{I} = [a,b]$ equals

$$(a,b) = \Big(\max\{\mu', m(\nu)\},\, \min\{\mu, M(\nu')\}\Big) = (\mu',\mu) \cap (m(\nu), M(\nu'))$$

and does not contain neither $m(\nu)$ nor $M(\nu')$. By Lemma 5.16, the functions $\mathcal{L}^{<}_{\inf}(\,\cdot\,,\nu)$ and $\mathcal{L}^{\leq}_{\inf}(\,\cdot\,,\nu)$ coincide on $\mathbb{R}\backslash\{m(\nu)\}$. It may be similarly shown that $\mathcal{L}^{>}_{\inf}(\,\cdot\,,\nu')$ and $\mathcal{L}^{\geq}_{\inf}(\,\cdot\,,\nu')$ coincide on $\mathbb{R}\backslash\{M(\nu')\}$. In particular, the functions $\mathcal{L}^{\leq}_{\inf}(\,\cdot\,,\nu)+\mathcal{L}^{\geq}_{\inf}(\,\cdot\,,\nu')$ and $\mathcal{L}^{<}_{\inf}(\,\cdot\,,\nu)+\mathcal{L}^{>}_{\inf}(\,\cdot\,,\nu')$ coincide on the interior of $\mathcal{I}$. Their infima over the interior of $\mathcal{I}$, which, by convexity, are equal to the infima over $\mathcal{I}$, are therefore equal.

Case 3: $\mathcal{I}$ is a singleton. This case arises if and only if $m(\nu)=M(\nu')$, as by definition, $m(\nu)\leq\mu$ and $M(\nu')\geq\mu'$. We then have $\mathcal{I}=\{m(\nu)\}=\{M(\nu')\}$, and both infima are equal to the values of the sums at $m(\nu)=M(\nu')$. By Lemma 5.16 and by symmetric results for $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$, on the one hand,

$$\mathcal{L}^{<}_{\inf}(m(\nu),\nu)=\mathcal{L}^{>}_{\inf}(M(\nu'),\nu')=+\infty\,,$$

and on the other hand,

$$\mathcal{L}^{\leq}_{\inf}(m(\nu),\nu)+\mathcal{L}^{\geq}_{\inf}(M(\nu'),\nu')=\log\frac{1}{\nu\{m(\nu)\}}+\log\frac{1}{\nu'\{M(\nu')\}}\,.$$

We get the desired equality if and only if either $\nu\{m(\nu)\}=0$ or $\nu\{M(\nu')\}=0$. □

5.5.2. Reminder: the Cramér-Chernoff Bound

In this section, we recall the statement of the highly classical Cramér-Chernoff bound: with the notation introduced in Section 5.3, for an N–sample $X_1,\dots,X_N$, distributed according to ν and of average denoted by $\overline{X}_N$,

$$\forall x\leq\mathrm{E}(\nu),\qquad \mathbb{P}(\overline{X}_N\leq x)\leq\exp(-N\,\phi^\star_\nu(x))\,, \tag{5.29}$$

and

$$\forall x\geq\mathrm{E}(\nu),\qquad \mathbb{P}(\overline{X}_N\geq x)\leq\exp(-N\,\phi^\star_\nu(x))\,. \tag{5.30}$$

Such a classical result would in principle not require to be proved here. However, it turns out that we will re-use parts of this proof in later proofs, like the application 5.31 of Jensen's inequality or the variations of $\phi^\star_\nu$ discussed at the end of this section. This is why, despite all, we now prove (5.29)–(5.30).

Proof. For all $\lambda<0$, by Markov's inequality first and then by independence,

$$\begin{aligned}\mathbb{P}(\overline{X}_N\leq x)=\mathbb{P}\Big(\mathrm{e}^{\lambda\overline{X}_N}\geq\mathrm{e}^{\lambda x}\Big)&\leq\mathrm{e}^{-\lambda x}\,\mathbb{E}\Big[\mathrm{e}^{\lambda\overline{X}_N}\Big]=\mathrm{e}^{-\lambda x}\,\Big(\mathbb{E}\big[\mathrm{e}^{\lambda X_1/N}\big]\Big)^N\\&=\exp(-\lambda x+N\,\phi_\nu(\lambda/N))=\exp\Big(-N\big(\lambda' x-\phi_\nu(\lambda')\big)\Big)\,,\end{aligned}$$

where $\lambda'=\lambda/N$. The bound also holds for $\lambda=\lambda'=0$ given that $\phi_\nu(0)=0$. Optimizing over $\lambda\leq0$ (or, equivalently, over $\lambda'\leq0$), we proved so far

$$\mathbb{P}(\overline{X}_N\leq x)\leq\exp\Big(-N\sup_{\lambda\leq0}\{\lambda x-\phi_\nu(\lambda)\}\Big)\,.$$

Now, by Jensen's inequality,

$$\forall\lambda\in\mathbb{R},\qquad \phi_\nu(\lambda)=\log\mathbb{E}\big[\mathrm{e}^{\lambda X}\big]\geq\lambda\,\mathbb{E}[X]=\lambda\,\mathrm{E}(\nu)\,; \tag{5.31}$$

therefore, for $x\leq\mathrm{E}(\nu)$,

$$\forall\lambda\geq0,\qquad \lambda x-\phi_\nu(\lambda)\leq\lambda\big(x-\mathrm{E}(\nu)\big)\leq0\,.$$

In particular,

$$0=-\phi_\nu(0)\leq\sup_{\lambda\leq0}\{\lambda x-\phi_\nu(\lambda)\}=\sup_{\lambda\in\mathbb{R}}\{\lambda x-\phi_\nu(\lambda)\}\stackrel{\text{def}}{=}\phi^\star_\nu(x)\,. \tag{5.32}$$

This concludes the proof of (5.29). The bound (5.30) follows by symmetry. □

We also note, in passing, that Jensen's inequality entails, for $x = \mathrm{E}(\nu)$, that

$$\forall \lambda \in \mathbb{R}, \qquad \lambda \mathrm{E}(\nu) - \phi_\nu(\lambda) \leq \lambda(\mathrm{E}(\nu) - \mathrm{E}(\nu)) = 0\,,$$

thus showing that $\phi_\nu^\star(\mathrm{E}(\nu)) = 0$. The property (5.32) and its counterpart for $x \geq \mathrm{E}(\nu)$ and $\lambda \geq 0$ actually show that $\phi_\nu^\star$ is non-increasing on $(-\infty, \mathrm{E}(\nu)]$ and non-decreasing on $[\mathrm{E}(\nu), +\infty)$.

5.5.3. Proofs and Details for Section 5.4.2: Rewriting of Φ as $\mathcal{L}$

We use the notation of Sections 5.2.1 and 5.3 and discuss conditions on models guaranteeing that $\Phi = \mathcal{L}$, i.e., that (5.22) holds. We do so for $\mathcal{D} = \mathcal{P}[0,1]$ in and for canonical one-parameter exponential families. Based on these two examples, we provide a set of conditions for general models at the end of the section A building block of these results is that for all these models $\mathcal{D}$, the functions $\mathcal{L}_{\inf}^{\leq}(\,\cdot\,, \nu)$ and $\mathcal{L}_{\inf}^{\geq}(\,\cdot\,, \nu)$ dominate the Fenchel-Legendre transform $\phi_\nu^\star$ defined in (5.15); we prove first.

All proofs of this section are immediate adaptations of a rather standard result, stated, among others, but in a slightly different form (and for the model $\mathcal{D}$ of all real-valued distributions with a first moment), by Boucheron et al. (2013, Exercise 4.13).

Remark. This rewriting of $\mathcal{L}_{\inf}^{\leq}(\,\cdot\,, \nu)$ or $\mathcal{L}_{\inf}^{\geq}(\,\cdot\,, \nu)$ as $\phi_\nu^\star$ claimed, e.g., by Lemma 5.5, can be seen as a counterpart to a similar rewriting of the $\mathcal{K}_{\inf}$ as the supremum of a function of $\lambda \in [0,1]$. More precisely, we recall (see Remark 1) that the $\mathcal{K}_{\inf}$ function is defined, for $\nu \in \mathcal{P}[0,1]$ and $x \in [0,1]$, as

$$\mathcal{K}_{\inf}(\nu, x) = \inf\{\mathrm{KL}(\nu, \zeta) : \ \zeta \in \mathcal{P}[0,1] \ \text{ s.t. } \ \mathrm{E}(\zeta) > x\}\,,$$

and Honda and Takemura (2015, Theorem 2) —see also Garivier et al., 2022, Lemma 18— show that

$$\mathcal{K}_{\inf}(\nu, x) = \sup_{0 \leq \lambda \leq 1} \mathbb{E}\left[\log\left(1 - \lambda \frac{X - x}{1 - x}\right)\right],$$

where X is a random variable distributed according to ν. In both cases, for $\mathcal{L}_{\inf}^{\leq}(\,\cdot\,, \nu)$ or $\mathcal{L}_{\inf}^{\geq}(\,\cdot\,, \nu)$, and for $\mathcal{K}_{\inf}$, being able to rewrite the infimum of Kullback-Leibler divergences as a supremum is not unexpected: a given Kullback-Leibler divergence can be formulated as a supremum, see (5.33), and equalities between $\inf \sup$ and $\sup \inf$ holds under suitable assumptions (provided, e.g., by Sion's lemma).

$\mathcal{L}_{\inf}^{\leq}(\,\cdot\,, \nu)$ and $\mathcal{L}_{\inf}^{\geq}(\,\cdot\,, \nu)$ dominate $\phi_\nu^\star$

This domination is a consequence of a variational formula (5.33) for the Kullback-Leibler divergences.

Lemma 5.18. *For all models $\mathcal{D}$ containing distributions with finite first moments, for all distributions $\nu \in \mathcal{D}$,*

$$\forall x \leq \mathrm{E}(\nu), \quad \phi_\nu^\star(x) \leq \mathcal{L}_{\inf}^{\leq}(x, \nu) \qquad \textit{and} \qquad \forall x \geq \mathrm{E}(\nu), \quad \phi_\nu^\star(x) \leq \mathcal{L}_{\inf}^{\geq}(x, \nu)\,.$$

Proof. We rely on a key variational formula for the Kullback-Leibler divergence, see Boucheron et al. (2013, Corollary 4.15): for all distributions ν, ν' over $\mathbb{R}$,

$$\begin{aligned} \mathrm{KL}(\nu', \nu) &= \sup\Big\{\mathbb{E}_{\nu'}[Y] - \log \mathbb{E}_\nu[\,\mathrm{e}^Y\,] : \ \text{r.v. } Y \in \mathbb{L}^1(\nu') \ \text{ s.t. } \ \mathbb{E}_\nu[\,\mathrm{e}^Y\,] < +\infty\Big\}\,, \\ &= \sup\Big\{\mathbb{E}_{\nu'}[Y] - \log \mathbb{E}_\nu[\,\mathrm{e}^Y\,] : \ \text{r.v. } Y \in \mathbb{L}^1(\nu')\Big\}\,, \end{aligned} \tag{5.33}$$

where the supremum is over random variables $Y : \mathbb{R} \to \mathbb{R}$ with a finite first moment with respect to ν', and where $\mathbb{E}_\nu$ and $\mathbb{E}_{\nu'}$ indicate that expectations are relative to ν and ν', respectively. In particular, when ν and ν' lie in $\mathcal{D}$, they admit finite first moments, hence all random variables of the form $Y = \lambda \, \mathrm{id}_{\mathbb{R}}$ are ν'–integrable, where $\mathrm{id}_{\mathbb{R}}$ denotes the identity function over $\mathbb{R}$ and where $\lambda \in \mathbb{R}$. We have $\mathbb{E}_{\nu'}[Y] = \lambda \, \mathrm{E}(\nu')$. A consequence of (5.33) and of the definition (5.15) of $\phi_\nu^\star$ is therefore that

$$\mathrm{KL}(\nu', \nu) \geq \sup_{\lambda \in \mathbb{R}} \Big\{ \lambda \, \mathrm{E}(\nu') - \log \mathbb{E}_\nu \big[\, \mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}} \, \big] \Big\} = \phi_\nu^\star(\mathrm{E}(\nu')) \, . \tag{5.34}$$

Using the variations of $\phi_\nu^\star$ indicated at the end of Section 5.5.2, we see that

$$\phi_\nu^\star(\mathrm{E}(\nu')) \geq \phi_\nu^\star(x) \quad \text{when} \quad \mathrm{E}(\nu') \leq x \leq \mathrm{E}(\nu) \quad \text{or} \quad \mathrm{E}(\nu') \geq x \geq \mathrm{E}(\nu) \, .$$

Therefore, taking an infimum in (5.34) yields, when $x \leq \mathrm{E}(\nu)$,

$$\mathcal{L}_{\mathrm{inf}}^{\leq}(x, \nu) = \inf \{ \mathrm{KL}(\nu', \nu) : \mathrm{E}(\nu') \leq x \} \geq \phi_\nu^\star(x) \, ,$$

and similarly for the other claimed inequality. □

The case of $\mathcal{P}[0,1]$

In this section, we focus on the model $\mathcal{P}[0,1]$ and prove that the inequalities of Lemma 5.18 are in fact equalities, as claimed by Lemma 5.5. This yields, in particular, the target equality (5.22), as discussed after the statement of Lemma 5.5.

Before proving Lemma 5.5, note that it holds for all $x \in \mathbb{R}$, that is, even outside of the $[0,1]$ interval, though the proof reveals that when x is smaller than the lower end $m(\nu)$ of the support of ν, we actually have $\phi_\nu^\star(x) = \mathcal{L}_{\mathrm{inf}}^{\leq}(x, \nu) = +\infty$. The counterpart statement $\phi_\nu^\star(x) = \mathcal{L}_{\mathrm{inf}}^{\geq}(x, \nu) = +\infty$ holds for x larger than the upper end $M(\nu)$ of the support of ν. The pieces of notation $m(\nu)$ and $M(\nu)$ were formally defined in Section 5.5.1.

Proof of Lemma 5.5. Note first that by Lemma 5.18, it suffices to prove that

$$\forall x \leq \mathrm{E}(\nu), \quad \phi_\nu^\star(x) \geq \mathcal{L}_{\mathrm{inf}}^{\leq}(x, \nu) \qquad \text{and} \qquad \forall x \geq \mathrm{E}(\nu), \quad \phi_\nu^\star(x) \geq \mathcal{L}_{\mathrm{inf}}^{\geq}(x, \nu) \, .$$

We only deal with the first inequality, namely $\mathcal{L}_{\mathrm{inf}}^{\leq}(x, \nu) \leq \phi_\nu^\star(x)$ for $x \leq \mathrm{E}(\nu)$, as the other one may be obtained by symmetric arguments.

In the case $x = \mathrm{E}(\nu)$, we have $\phi_\nu^\star(\mathrm{E}(\nu)) = 0$, as stated at the end of Section 5.5.2, and $\mathcal{L}_{\mathrm{inf}}^{\leq}(\mathrm{E}(\nu), \nu) = 0$, as can be seen by taking $\zeta = \nu$ in the infimum defining $\mathcal{L}_{\mathrm{inf}}^{\leq}$. We therefore only consider $x < \mathrm{E}(\nu)$ in the sequel. We will rely on the standard fact that, by Hölder's inequality, the logarithmic moment-generating function

$$\phi_\nu : \lambda \in \mathbb{R} \longmapsto \log \mathbb{E}_\nu \big[\, \mathrm{e}^{\lambda \mathrm{id}_{[0,1]}} \, \big] \, ,$$

is convex, where $\mathrm{id}_{[0,1]}$ denotes the identity function on $[0,1]$. Also, by two applications of a standard theorem of differentiation under the integral, given that ν is supported by $[0,1]$, we have that ϕ_ν is continuously differentiable over $\mathbb{R}$, with derivative

$$\phi_\nu' : \lambda \in \mathbb{R} \longmapsto \frac{\mathbb{E}_\nu \big[\mathrm{id}_{[0,1]} \, \mathrm{e}^{\lambda \mathrm{id}_{[0,1]}} \big]}{\mathbb{E}_\nu \big[\, \mathrm{e}^{\lambda \mathrm{id}_{[0,1]}} \, \big]} \, .$$

By convexity of ϕ_ν, this derivative is non-decreasing. Therefore, the limit of ϕ_ν' at $-\infty$ exists; we denote it by ℓ and have that a priori $\ell \in \{-\infty\} \cup \mathbb{R}$. We now prove that actually,

$$\ell \stackrel{\mathrm{def}}{=} \lim_{\lambda \to -\infty} \phi_\nu'(\lambda) = m(\nu) \, . \tag{5.35}$$

On the one hand, by definition of $m(\nu)$, we have $\mathrm{id}_{[0,1]} \geq m(\nu)$ ν-a.s., which entails $\phi_\nu'(\lambda) \geq m(\nu)$ for all $\lambda \in \mathbb{R}$, and hence, $\ell \geq m(\nu)$. On the other hand, as ϕ_ν' is non-decreasing, it is always larger than its limit ℓ at $-\infty$:

$$\forall \lambda \in \mathbb{R}, \qquad \phi_\nu'(\lambda) \geq \ell\,, \qquad \text{thus,} \qquad \mathbb{E}_\nu\Big[\big(\mathrm{id}_{[0,1]} - \ell\big)\ \mathrm{e}^{\lambda \mathrm{id}_{[0,1]}}\Big] \geq 0\,, \tag{5.36}$$

$$\text{or} \qquad \mathbb{E}_\nu\Big[\big(\mathrm{id}_{[0,1]} - \ell\big)\ \mathrm{e}^{\lambda(\mathrm{id}_{[0,1]} - \ell)}\Big] \geq 0\,. \tag{5.37}$$

The last inequality and limit arguments as $\lambda \to -\infty$ impose that $\mathrm{id}_{[0,1]} - \ell \geq 0$ ν-a.s., which in turn entails that $\ell \leq m(\nu)$. This concludes the proof of (5.35).

The various properties exhibited above for ϕ_ν, including the fact that the derivative ϕ_ν' takes values in $\big[m(\nu), +\infty\big)$, entail that the function

$$\Lambda : \lambda \in \mathbb{R} \longmapsto \lambda x - \phi_\nu(\lambda)$$

is concave, continuously differentiable, with a non-increasing derivative Λ' taking values in the interval $\big(-\infty, x - m(\nu)\big]$ and with limit $x - m(\nu)$ at $-\infty$.

We split the analysis of the case $x < \mathrm{E}(\nu)$ into three sub-cases, depending on the respective positions of x and $m(\nu)$, and recall that we want to show that $\mathcal{L}^{\leq}_{\inf}(x, \nu) \leq \phi^\star_\nu(x)$.

Case 1: $x > m(\nu)$. By Jensen's inequality (5.31) and given that we consider $x < \mathrm{E}(\nu)$, the limit of Λ at $+\infty$ equals $-\infty$. The limit of Λ at $-\infty$ also equals $-\infty$, as the derivative Λ' has limit $x - m(\nu) > 0$ at $-\infty$. By concavity of Λ and the fact that Λ' is continuous, this implies the existence of some $\lambda^\star \in \mathbb{R}$ such that

$$\Lambda'(\lambda^\star) = x - \phi_\nu'(\lambda^\star) = 0 \qquad \text{and} \qquad \phi^\star_\nu(x) = \sup_{\lambda \in \mathbb{R}} \big\{\Lambda(\lambda)\big\} = \Lambda(\lambda^\star)\,.$$

Denoting by $\zeta_{\lambda^\star}$ the distribution absolutely continuous with respect to ν with density

$$\frac{\mathrm{d}\zeta_{\lambda^\star}}{\mathrm{d}\nu} = \frac{\mathrm{e}^{\lambda^\star \mathrm{id}_{[0,1]}}}{\mathbb{E}_\nu\big[\,\mathrm{e}^{\lambda^\star \mathrm{id}_{[0,1]}}\,\big]} = \mathrm{e}^{\lambda^\star \mathrm{id}_{[0,1]} - \phi_\nu(\lambda^\star)}\,,$$

we have $\mathbb{E}_{\zeta_{\lambda^\star}}\big[\mathrm{id}_{[0,1]}\big] = \mathrm{E}(\zeta_{\lambda^\star}) = \phi_\nu'(\lambda^\star) = x$. Therefore, by definition of $\mathcal{L}^{\leq}_{\inf}(x, \nu)$ and of the Kullback-Leibler divergence,

$$\mathcal{L}^{\leq}_{\inf}(x, \nu) \leq \mathrm{KL}(\zeta_{\lambda^\star}, \nu) = \mathbb{E}_{\zeta_{\lambda^\star}}\left[\log \frac{\mathrm{d}\zeta_{\lambda^\star}}{\mathrm{d}\nu}\right] = \lambda^\star\, \mathbb{E}_{\zeta_{\lambda^\star}}\big[\mathrm{id}_{[0,1]}\big] - \phi_\nu(\lambda^\star) = \Lambda(\lambda^\star) = \phi^\star_\nu(x)\,.$$

Case 2: $x = m(\nu)$. In that case, $\Lambda' \to 0$ at $-\infty$ and Λ' is non-increasing, thus $\Lambda' \leq 0$ on $\mathbb{R}$ and Λ is non-increasing on $\mathbb{R}$. Thus,

$$\phi^\star_\nu(m(\nu)) = \sup_{\lambda \in \mathbb{R}} \big\{\Lambda(\lambda)\big\} = \lim_{\lambda \to -\infty} \Lambda(\lambda) = \lim_{\lambda \to -\infty} -\log \mathbb{E}_\nu\Big[\mathrm{e}^{\lambda(\mathrm{id}_{[0,1]} - m(\nu))}\Big]\,.$$

By monotone convergence based on $\mathrm{id}_{[0,1]} - m(\nu) \geq 0$ ν-a.s.,

$$\lim_{\lambda \to -\infty} -\log \mathbb{E}_\nu\Big[\mathrm{e}^{\lambda(\mathrm{id}_{[0,1]} - m(\nu))}\Big] = -\log \nu\big\{m(\nu)\big\}\,,$$

whether $\nu\big\{m(\nu)\big\}$ is positive or null. Moreover, Lemma 5.16 states that

$$\mathcal{L}^{\leq}_{\inf}(m(\nu), \nu) = -\log \nu\big\{m(\nu)\big\}\,.$$

We therefore have $\mathcal{L}^{\leq}_{\inf}(x, \nu) = \phi^\star_\nu(x)$ in this case.

Case 3: $x < m(\nu)$. In that case, as $\Lambda' \to x - m(\nu) < 0$ at $-\infty$, we get that $\Lambda \to +\infty$ at $-\infty$, thus $\phi^\star_\nu(x) = \sup \Lambda = +\infty$. Now, no distribution $\zeta \in \mathcal{P}[0,1]$ with $\mathrm{E}(\zeta) \leq x$, if some exists, can be absolutely continuous with respect to ν; indeed, $x < m(\nu)$ imposes that ζ puts some probability mass to the left of the support of ν. Therefore, $\mathrm{KL}(\zeta, \nu) = +\infty$. All in all, $\mathcal{L}^{\leq}_{\inf}(x, \nu)$ appears as the infimum of either an empty set or of $+\infty$ values, so that $\mathcal{L}^{\leq}_{\inf}(x, \nu) = +\infty$. In this case as well, $\mathcal{L}^{\leq}_{\inf}(x, \nu) = \phi^\star_\nu(x)$, both being equal to $+\infty$. □

The case of canonical one-parameter exponential models $\mathcal{D}_{\exp}$

In this section, we show that the target equality (5.22) is satisfied by so-called canonical one-parameter exponential families $\mathcal{D}_{\exp}$. We recall that those models are defined at the beginning of Section 2.2.3 and follow the notation of that Section: ρ is the reference measure, b the normalizing function, Θ is the natural space parameter, and $\mathcal{M} = (\mu_-, \mu_+)$ is the open interval of the expectations of distributions in $\mathcal{D}_{\exp}$.

The mean-parametrized Kullback-Leibler divergence of the model is denoted by d and defined on $\mathcal{M} \times \mathcal{M}$, where. In the following, we extend d to $\mathbb{R} \times \mathbb{R}$ by $+\infty$ values outside of $\mathcal{M} \times \mathcal{M}$.

A direct application of the continuity and monotonicity properties of d is that all functions $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$, $\mathcal{L}^{\geq}_{\inf}$ coincide with d in the sense of the stated equalities (5.12) and (5.13). Indeed and for instance, we have, for $\nu \in \mathcal{D}_{\exp}$ and $x \leq \mathrm{E}(\nu)$ with $x \in \mathcal{M}$:

$$\mathcal{L}^{<}_{\inf}(x, \nu) = \inf_{\mu < x} \{d(\mu, \nu)\} = \lim_{\substack{\mu \to x \\ \mu < x}} d(\mu, \nu) = d(x, \nu)\,.$$

When $x \notin \mathcal{M}$, by the convention on the infimum of an empty set, $\mathcal{L}^{<}_{\inf}(x, \nu) = +\infty$, while by our definition of d outside $\mathcal{M} \times \mathcal{M}$, we also have $d(x, \nu) = +\infty$. But as Lemma 5.20 below illustrates, we will only be interested in the behaviors on $\mathcal{M} \times \mathcal{M}$.

We now state a monotonicity property of the Chernoff-information-type quantity L defined for exponential models in (5.14). This property was referred to in Example 5.1, when indicating that arms can be equivalently ranked in descending expectations or ascending values of $L(\,\cdot\,, \mu^\star)$.

Lemma 5.19. *Consider a canonical one-parameter exponential family $\mathcal{D}_{\exp}$ and fix any $\mu \in \mathcal{M}$. Then $L(\,\cdot\,, \mu)$ is non-increasing on $(\mu_-, \mu]$.*

Proof. Fix $\mu_- < \mu_2 \leq \mu_1 \leq \mu$. To get the desired inequality $L(\mu_2, \mu) \geq L(\mu_1, \mu)$, it suffices to show, by (5.14), that

$$\forall y \in [\mu_2, \mu], \qquad d(y, \mu_2) + d(y, \mu) \geq \min_{x \in [\mu_1, \mu]} d(x, \mu_1) + d(x, \mu) \overset{\text{def}}{=} L(\mu_1, \mu)\,. \tag{5.38}$$

We distinguish two cases. If $\mu_2 \leq \mu_1 \leq y \leq \mu$, then, since $d(y, \,\cdot\,)$ is increasing on $(\mu_-, y]$, we have $d(y, \mu_2) \geq d(y, \mu_1)$, from which the inequality (5.38) follows by considering $x = y$. If $\mu_2 \leq y \leq \mu_1 \leq \mu$, then similarly $d(y, \mu) \geq d(\mu_1, \mu)$, which yields

$$\underbrace{d(y, \mu_2)}_{\geq 0} + d(y, \mu) \geq d(\mu_1, \mu) = \underbrace{d(\mu_1, \mu_1)}_{=0} + d(\mu_1, \mu)\,,$$

from which the inequality (5.38) follows by considering $x = \mu_1$. □

A slightly weaker version of Lemma 5.5, sufficient for our purposes. We may now come back to the proof of the target equality (5.22) for canonical one-parameter exponential families. The following slightly weaker version of Lemma 5.5 is enough to yield (5.22), given the rewritings (5.12) and (5.13).

Lemma 5.20. *Consider a canonical one-parameter exponential family $\mathcal{D} = \mathcal{D}_{\exp}$. For all $\nu \in \mathcal{D}_{\exp}$,*

$$\forall x \in \mathcal{M}, \qquad \phi^\star_\nu(x) = d(x, \mathrm{E}(\nu))\,.$$

The result of the lemma holds, by conventions, for $x < \mu_-$ or $x > \mu_+$, but does not hold in general for $x \in \{\mu_-, \mu_+\}$.

Proof. By Lemma 5.18, we only need to show that $\phi_\nu^\star(x) \geq d(x, \mathrm{E}(\nu))$. Given the definition (5.15) of $\phi_\nu^\star$ as a supremum, it suffices to exhibit a $\lambda^\star \in \mathbb{R}$ such that

$$d(x, \mathrm{E}(\nu)) = \lambda^\star x - \phi_\nu(\lambda^\star)\,. \tag{5.39}$$

Let $\theta_1 \in \Theta$ be such that $\nu = \nu_{\theta_1}$ and $\theta_2 = (b')^{-1}(x) \in \Theta$ be such that $\mathrm{E}(\nu_{\theta_2}) = x$. We will prove (5.39) with $\lambda^\star = \theta_2 - \theta_1$. Given the closed-form expression of the densities (2.15), the distribution ν_{θ_2} is absolutely continuous with respect to ν_{θ_1}, with density given by $(\theta_2 - \theta_1)\mathrm{id}_{\mathbb{R}} - \big(b(\theta_2) - b(\theta_1)\big)$. Therefore, by definition of the Kullback-Leibler divergence,

$$\begin{aligned} d(x, \mathrm{E}(\nu)) = \mathrm{KL}(\nu_{\theta_2}, \nu_{\theta_1}) &= \mathbb{E}_{\nu_{\theta_2}}\left[\log \frac{\mathrm{d}\nu_{\theta_2}}{\mathrm{d}\nu_{\theta_1}}\right] = \mathbb{E}_{\nu_{\theta_2}}\Big[(\theta_2 - \theta_1)\,\mathrm{id}_{\mathbb{R}} - \big(b(\theta_2) - b(\theta_1)\big)\Big] \\ &= (\theta_2 - \theta_1)\,\mathrm{E}(\nu_{\theta_2}) - \big(b(\theta_2) - b(\theta_1)\big) = \lambda^\star x - \big(b(\theta_2) - b(\theta_1)\big)\,. \end{aligned} \tag{5.40}$$

To obtain (5.39), it only remains to show that $b(\theta_2) - b(\theta_1) = \phi_\nu(\lambda^\star)$. Using the closed-form expressions (2.16) of b at θ_2 and (2.15) of the density at θ_1, we obtain

$$\begin{aligned} b(\theta_2) = \log \int_{\mathbb{R}} \mathrm{e}^{\theta_2 y}\,\mathrm{d}\rho(y) &= b(\theta_1) + \log \int_{\mathbb{R}} \mathrm{e}^{(\theta_2 - \theta_1) y} \overbrace{e^{\theta_1 y - b(\theta_1)}\,\mathrm{d}\rho(y)}^{=\mathrm{d}\nu_{\theta_1}(y) = \mathrm{d}\nu(y)} \\ &= b(\theta_1) + \log \int_{\mathbb{R}} \mathrm{e}^{\lambda^\star y}\,\mathrm{d}\nu(y) = b(\theta_1) + \phi_\nu(\lambda^\star)\,, \end{aligned} \tag{5.41}$$

which concludes the proof. □

Remark. A more direct approach bypassing Lemma 5.18 can be followed with $\mathcal{D}_{\exp}$ models, along the following lines. The result (5.41) can be generalized into

$$\forall \theta \in \Theta, \qquad \phi_\nu(\theta - \theta_1) = b(\theta) - b(\theta_1)\,. \tag{5.42}$$

As b is differentiable on Θ, the function ϕ_ν is also differentiable; at $\lambda^\star = \theta_2 - \theta_1$, we have

$$\phi_\nu'(\lambda^\star) = \phi_\nu'(\theta_2 - \theta_1) = b'(\theta_2) = x\,.$$

Thus, the derivative of the strictly concave function $\Lambda : \lambda \in \mathbb{R} \longmapsto \lambda x - \phi_\nu(\lambda)$ vanishes at $\lambda^\star$, which is therefore the argument of its maximum: $\phi_\nu^\star(x) = \Lambda(\lambda^\star)$. The closed-form calculation (5.40) and the rewriting (5.42) then lead to Lemma 5.20.

Conditions for general models

In this section, we extend Lemma 5.5, and thus the target equality (5.22), to more general models. We did so by mimicking the proof of Lemma 5.5: the result below can certainly be improved. We extend as follows the definitions of the lower and upper ends $m(\nu)$ and $M(\nu)$ of the closed support $\mathrm{Supp}(\nu)$ of a distribution ν over $\mathbb{R}$:

$$m(\nu) = \inf\big(\mathrm{Supp}(\nu)\big) \in \mathbb{R} \cup \{-\infty\} \qquad \text{and} \qquad M(\nu) = \sup\big(\mathrm{Supp}(\nu)\big) \in \mathbb{R} \cup \{+\infty\}\,.$$

Lemma 5.21. *Consider a model $\mathcal{D}$ containing distributions ν over $\mathbb{R}$ with finite first moments and with exponential moments: $\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}} \in \mathbb{L}^1(\nu)$ for all $\lambda \in \mathbb{R}$. Assume that the model $\mathcal{D}$ is stable by exponential reweighting of densities: for all $\nu \in \mathcal{D}$, for all $\lambda \in \mathbb{R}$, the distribution ν_λ with density*

$$\frac{\mathrm{d}\nu_\lambda}{\mathrm{d}\nu} = \frac{\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}}{\mathbb{E}_\nu\big[\,\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}\,\big]} \qquad \textit{with respect to } \nu \tag{5.43}$$

also belongs to $\mathcal{D}$. Assume also that δ_x, the Dirac mass at x, belongs to $\mathcal{D}$ whenever there exists $\nu \in \mathcal{D}$ with $x \in \{m(\nu), M(\nu)\} \cap \mathbb{R}$ and $\nu\{x\} > 0$; put differently, if a distribution $\nu \in \mathcal{D}$ puts some probability mass on an end x of its closed support, then the Dirac mass at x belongs to $\mathcal{D}$.

Then, for all $\nu \in \mathcal{D}$,

$$\forall x \le \mathrm{E}(\nu), \quad \phi_\nu^\star(x) = \mathcal{L}^{\le}_{\inf}(x,\nu) \qquad \textit{and} \qquad \forall x \ge \mathrm{E}(\nu), \quad \phi_\nu^\star(x) = \mathcal{L}^{\ge}_{\inf}(x,\nu)\,.$$

Proof. By symmetry and by Lemma 5.18, we only need to prove that

$$\forall x \le \mathrm{E}(\nu), \quad \phi_\nu^\star(x) \ge \mathcal{L}^{\le}_{\inf}(x,\nu)\,.$$

For $x = \mathrm{E}(\nu)$, we have $\phi_\nu^\star(\mathrm{E}(\nu)) = 0 = \mathcal{L}^{\le}_{\inf}(\mathrm{E}(\nu),\nu)$, as stated at the end of Section 5.5.2 and by taking $\zeta = \nu$ in the infimum defining $\mathcal{L}^{\le}_{\inf}$, respectively. Before moving to the case $x < \mathrm{E}(\nu)$, we establish a few properties of ϕ_ν based on the assumptions of Lemma 5.21. All random variables $\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}$ are ν–integrable, for $\lambda \in \mathbb{R}$, which entails, by application of a standard theorem of differentiation under the integral sign together with local domination arguments of the form

$$\forall \lambda \in (\lambda_-, \lambda_+), \qquad \big|\mathrm{id}_{\mathbb{R}}\ \mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}\big| \le \big|\mathrm{id}_{\mathbb{R}}\big|\,\big(\mathrm{e}^{\lambda_- \mathrm{id}_{\mathbb{R}}} + \mathrm{e}^{\lambda_+ \mathrm{id}_{\mathbb{R}}}\big) \le \big(\mathrm{e}^{\mathrm{id}_{\mathbb{R}}} + \mathrm{e}^{-\mathrm{id}_{\mathbb{R}}}\big)\big(\mathrm{e}^{\lambda_- \mathrm{id}_{\mathbb{R}}} + \mathrm{e}^{\lambda_+ \mathrm{id}_{\mathbb{R}}}\big)\,,$$

that ϕ_ν is differentiable over $\mathbb{R}$, with derivative given by

$$\phi_\nu' : \lambda \in \mathbb{R} \longmapsto \frac{\mathbb{E}_\nu\big[\mathrm{id}_{\mathbb{R}}\ \mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}\big]}{\mathbb{E}_\nu\big[\,\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}\,\big]}\,. \tag{5.44}$$

Hölder's inequality still entails that ϕ_ν is convex, thus its derivative ϕ_ν' is non-decreasing; therefore, ϕ_ν' admits a limit $\ell \in \{-\infty\} \cup \mathbb{R}$ at $-\infty$. Actually, we have $\ell = m(\nu)$, as can be seen by combining the following facts. First, by definition, $\mathrm{id}_{\mathbb{R}} \ge m(\nu)$ ν-a.s., thus $\phi_\nu' \ge m(\nu)$, hence $\ell \ge m(\nu)$. As a consequence, if $\ell = -\infty$, then we also have $m(\nu) = -\infty$. Otherwise, if $\ell \in \mathbb{R}$, the same arguments as in (5.36)–(5.37) show that $\mathrm{id}_{\mathbb{R}} - \ell \ge 0$ ν-a.s., i.e., $\ell \le m(\nu)$.

We may now come back to establishing $\phi_\nu^\star(x) \ge \mathcal{L}^{\le}_{\inf}(x,\nu)$ in the case $x < \mathrm{E}(\nu)$. We consider three sub-cases, depending on the respective positions of x and $m(\nu)$.

Case 1: $x > m(\nu)$. The properties of ϕ_ν ensure, exactly as in Case 1 of the proof of Lemma 5.5, the existence of $\lambda^\star$ such that $\phi_\nu'(\lambda^\star) = x$ and $\phi_\nu^\star(x) = \lambda^\star x - \phi_\nu(\lambda^\star)$. Given the assumption (5.43), we may consider the distribution $\nu_{\lambda^\star} \in \mathcal{D}$. We note, again exactly as in Case 1 of the proof of Lemma 5.5 and given the closed-form expression (5.44) for ϕ_ν', that $\mathrm{E}(\nu_{\lambda^\star}) = \phi_\nu'(\lambda^\star)$, thus $\mathrm{E}(\nu_{\lambda^\star}) = x$. Finally, an explicit computation yields

$$\mathrm{KL}(\nu_{\lambda^\star}, \nu) = \lambda^\star \mathrm{E}(\nu_{\lambda^\star}) - \log \mathbb{E}_\nu\big[\,\mathrm{e}^{\lambda \mathrm{id}_{\mathbb{R}}}\,\big] = \lambda^\star x - \phi_\nu(\lambda^\star) = \phi_\nu^\star(x)\,.$$

By the defining infimum of $\mathcal{L}^{\le}_{\inf}(x,\nu)$, we have indeed $\mathcal{L}^{\le}_{\inf}(x,\nu) \le \mathrm{KL}(\nu_{\lambda^\star}, \nu) = \phi_\nu^\star(x)$.

Case 2: $x = m(\nu)$. In particular, $m(\nu) \in \mathbb{R}$, which allows us to follow the monotone-convergence arguments of Case 2 of the proof of Lemma 5.5 and get the equality $\phi_\nu^\star(m(\nu)) = -\log \nu\{m(\nu)\}$. Now, for the second part of this sub-case, we also adapt an argument of the second part of the proof

of Lemma 5.16 (in Section 5.5.1), namely, the fact that either there exists at most one distribution $\zeta \in \mathcal{D}$ absolutely continuous with respect to ν and satisfying $\mathrm{E}(\zeta) \leq m(\nu)$, namely, $\zeta = \delta_{m(\nu)}$, the Dirac mass at $m(\nu)$. The latter is indeed absolutely continuous with respect to ν if and only if $\nu\{m(\nu)\} > 0$. When $\nu\{m(\nu)\} > 0$, we have $\delta_{m(\nu)} \in \mathcal{D}$ by the Dirac assumption of the lemma, so that

$$\mathcal{L}^{\leq}_{\inf}(m(\nu), \nu) = \mathrm{KL}(\delta_{m(\nu)}, \nu) = -\log \nu\{m(\nu)\}\,.$$

Otherwise, when $\nu\{m(\nu)\} = 0$, the infimum defining $\mathcal{L}^{\leq}_{\inf}(m(\nu), \nu)$ is either over an empty set or of $+\infty$ values, and thus equals $+\infty = -\log \nu\{m(\nu)\}$. In both situations, we obtained $\mathcal{L}^{\leq}_{\inf}(m(\nu), \nu) = \phi^{\star}_{\nu}(m(\nu))$.

Case 3: $x < m(\nu)$. In particular, $m(\nu) \in \mathbb{R}$ in this sub-case as well, which allows us to repeat the exact same arguments as in Case 3 of the proof of Lemma 5.5: we may show that both $\mathcal{L}^{\leq}_{\inf}(x, \nu)$ and $\phi^{\star}_{\nu}(x)$ are equal to $+\infty$. □

5.5.4. Proof of the Normality of the Models $\mathcal{P}[0,1]$ and $\mathcal{D}_{\exp}$

In this section, we show that $\mathcal{P}[0,1]$ and canonical one-parameter exponential models are normal.

Proposition 5.22. *$\mathcal{P}[0,1]$ is a normal model.*

Proof. We fix $\nu \in \mathcal{P}[0,1]$, a real $x \geq \mathrm{E}(\nu)$, and $\varepsilon > 0$. Recall the piece of notation $M(\nu)$ for the upper end of the support of ν, as introduced in Section 5.5.1. As in Case 3 of the proof of Lemma 5.5, we note that when $x \geq M(\nu)$, there exists no distribution $\zeta \in \mathcal{P}[0,1]$ absolutely continuous with respect to ν and such that $\mathrm{E}(\zeta) > x$; hence, both infima in Definition 5.12 equal $+\infty$. We now tackle the case where $\mathrm{E}(\nu) \leq x < M(\nu)$. For all $\delta > 0$, we introduce

$$x'_{\delta} = \min\left\{x + \delta,\, \frac{x + M(\nu)}{2}\right\} < M(\nu)\,.$$

Case 1 of the proof of Lemma 5.5 and Lemma 5.18 reveal (by symmetry) that for each $\delta > 0$, there exists a distribution $\zeta_{\delta} \in \mathcal{P}[0,1]$ with expectation x'_{δ} and such that $\mathcal{L}^{\geq}_{\inf}(x'_{\delta}, \nu) = \phi^{\star}_{\nu}(x'_{\delta}) = \mathrm{KL}(\zeta_{\delta}, \nu)$. By Lemma 5.16, $\mathcal{L}^{\geq}_{\inf}(x'_{\delta}, \nu) = \mathcal{L}^{>}_{\inf}(x'_{\delta}, \nu)$ and $\mathcal{L}^{>}_{\inf}(\,\cdot\,, \nu)$ is continuous on $(-\infty, M(\nu))$. Putting all these elements together, we obtain

$$\begin{aligned}\mathcal{L}^{>}_{\inf}(x, \nu) = \lim_{\delta \to 0} \mathcal{L}^{>}_{\inf}(x'_{\delta}, \nu) = \liminf_{\delta \to 0} \mathrm{KL}(\zeta_{\delta}, \nu) &\geq \inf\{\mathrm{KL}(\zeta_{\delta}, \nu) :\ \delta \in (0, \varepsilon)\} \\ &\geq \inf\{\mathrm{KL}(\zeta, \nu) :\ \zeta \in \mathcal{D} \ \text{ s.t. } \ x + \varepsilon > \mathrm{E}(\zeta) > x\}\,,\end{aligned}$$

where the first inequality is by the very definition of a $\liminf$. □

Proposition 5.23. *All canonical one-parameter exponential models $\mathcal{D}_{\exp}$ are normal.*

Proof. The proof consists of rewriting $\mathcal{L}^{>}_{\inf}$ as d, as indicated by (5.13), and using the regularity properties for d exhibited in Section 5.5.3 (see page 151). We fix $\nu \in \mathcal{D}_{\exp}$, a real $x \geq \mathrm{E}(\nu)$, and $\varepsilon > 0$. When $x \geq M(\nu)$, the same argument as in the previous proposition shows that both infima equal $+\infty$. For $x < M(\nu)$, we introduce $\delta \in (0, \mu_+ - x)$ and write

$$\begin{aligned}\mathcal{L}^{>}_{\inf}(x, \nu) = d(x, \mathrm{E}(\nu)) = \lim_{\delta \to 0} d(x + \delta, \mathrm{E}(\nu)) &= \inf\Big\{d(x + \delta, \mathrm{E}(\nu)) : \delta \in (0, \varepsilon)\Big\} \\ &= \inf\{\mathrm{KL}(\zeta, \nu) :\ \zeta \in \mathcal{D} \ \text{ s.t. } \ x + \varepsilon > \mathrm{E}(\zeta) > x\}\,,\end{aligned}$$

where the second and third equalities follow, respectively, by continuity of $d(\,\cdot\,, \mathrm{E}(\nu))$ on $\mathcal{M}$ and by the fact that this function is non-decreasing on $(x, \mu_+) \subset [\mathrm{E}(\nu), \mu_+)$, and the final equality is by the rewriting (2.17). □

5.6. Additional Comments for the Literature Review

This section is devoted to additional discussions concerning the fixed-budget literature. More precisely, we discuss in detail two gap-based lower bounds that we believe are somewhat detached from the spirit of the chapter, namely, the minimax lower bound of Carpentier and Locatelli (2016) in Section 5.6.1 and the Bretagnolle-Huber technique in Section 5.6.2.

5.6.1. The Minimax Lower Bound of Carpentier and Locatelli (2016)

Carpentier and Locatelli (2016, Theorem 1) proved (slightly stronger versions of) the following (non-asymptotic) minimax lower bound. Consider the model $\mathcal{B}_{[1/4,\,3/4]}$ of Bernoulli distributions $\mathrm{Ber}(p)$ with parameters $p \in [1/4,\, 3/4]$. For all sequences of strategies that are consistent on $\mathcal{B}_{[1/4,\,3/4]}$, for all $T \geq 0.14\, K^4 \log(6KT)$,

$$\exists\, \underline{\nu} \text{ in } \mathcal{B}_{[1/4,\,3/4]}, \qquad \frac{1}{T} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) \geq -\frac{400}{\log K} \left(\sum_{a \neq a^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \right)^{-1} - \frac{\log 6}{T}\,, \tag{5.45}$$

where, of course, we may rather use the weaker lower bound based on

$$-\left(\sum_{a \neq a^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \right)^{-1} \geq - \min_{2 \leq k \leq K} \frac{\Delta_{(k)}^2}{k}\,.$$

However, the bound (5.45) is different in nature from the lower bounds considered in this chapter, as first and foremost, it only guarantees a $1/\log K$ improvement of the lower bound (5.8) of Audibert et al. (2010) for a single bandit problem $\underline{\nu}$ (actually belonging to a known collection of K bandit problems). This is in strong contrast with the uniform instance-dependent lower bounds presented in this chapter: bounds holding simultaneously for all bandit problems of a given model. Second, the proof of the result (see the simpler proof provided below for Proposition 5.24 stated next) is truly gap-based and does not seem to extend in any obvious way to non-parametric models.

As mentioned above, the proof of (5.45) in Carpentier and Locatelli (2016) uses only K different bandit problems in $\mathcal{B}_{[1/4,\,3/4]}$. We may therefore resort to the pigeonhole principle to exchange, in some sense, the "for all $T \geq 0.14K^4 \log(6KT)$" and "there exists $\underline{\nu}$ in $\mathcal{B}_{[1/4,\,3/4]}$" parts. More precisely, we obtain, from (5.45) the following proposition. For the sake of completeness, we provide a self-contained proof of this proposition closely following the original arguments by Carpentier and Locatelli (2016), except for the change-of-measure argument, for which we rather resort to Lemma 5.7. Doing so, we are able to improve the numerical factor 400 that would follow from (5.45) into a smaller factor of 30.

Proposition 5.24. *Fix $K \geq 3$ and consider the model $\mathcal{B}_{[1/4,\,3/4]}$ of Bernoulli distributions $\mathrm{Ber}(p)$ with parameters $p \in [1/4,\, 3/4]$. For all consistent sequences of strategies on $\mathcal{B}_{[1/4,\,3/4]}$, there exists an increasing sequence of budgets $(T_n)_{n \geq 1}$ such that*

$$\exists\, \underline{\nu} \text{ in } \mathcal{B}_{[1/4,\,3/4]}, \qquad \liminf_{n \to +\infty} \frac{1}{T_n} \log \mathbb{P}_{\underline{\nu}}(\hat{a}_{T_n} \neq a^\star(\underline{\nu})) \geq -\frac{30}{\log K} \left(\sum_{a \neq a^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \right)^{-1}. \tag{5.46}$$

Proof. We consider some base Bernoulli bandit problem $\underline{\nu}^{\mathsf{base}} = (\nu_1^{\mathsf{base}}, \dots, \nu_K^{\mathsf{base}})$, where

$$\nu_1^{\mathsf{base}} = \mathrm{Ber}(1/2) \qquad \text{and} \qquad \forall j \in \{2, \dots, K\}, \quad \nu_j^{\mathsf{base}} = \mathrm{Ber}(p_j)\,,$$

for parameters $p_j \in [1/4,\, 1/2)$ to be specified later. For each $k \in \{2, \dots, K\}$, we then define the alternative bandit problem $\underline{\nu}^{(k)} = (\nu_1^{(k)}, \dots, \nu_K^{(k)})$ as follows:

$$\nu_j^{(k)} = \begin{cases} \mathrm{Ber}(1 - p_k) & \text{if } j = k, \\ \nu_j^{\mathsf{base}} & \text{if } j \neq k. \end{cases}$$

Given the constraints on the p_j, the unique optimal arm of $\underline{\nu}^{\mathsf{base}}$ is $a^\star(\underline{\nu}^{\mathsf{base}}) = 1$, while the unique optimal arm of $\underline{\nu}^{(k)}$ is $a^\star(\underline{\nu}^{(k)}) = k$. We introduce, for a given bandit problem $\underline{\nu}$

$$H_\Sigma(\underline{\nu}) \stackrel{\mathrm{def}}{=} \sum_{a \neq a^\star(\underline{\nu})} \frac{1}{\Delta_a^2}\,;$$

the right-hand side of (5.46) may be rewritten as $(34/\log K)\, H_\Sigma(\underline{\nu})^{-1}$. The sub-optimality gaps of the arms of $\underline{\nu}^{\mathsf{base}}$ equal $\Delta_j^{\mathsf{base}} = 1/2 - p_j$ for $j \neq 1$, while the ones of $\underline{\nu}^{(k)}$ equal

$$\begin{aligned} &\forall j \neq k, \qquad \Delta_j^{(k)} = 1 - p_k - p_j = (1/2 - p_k) + (1/2 - p_j) = \Delta_k^{\mathsf{base}} + \Delta_j^{\mathsf{base}}\,, \\ &\text{thus} \qquad H_\Sigma(\underline{\nu}^{(k)}) = \sum_{j \neq k} \frac{1}{\big(\Delta_k^{\mathsf{base}} + \Delta_j^{\mathsf{base}}\big)^2}\,. \end{aligned} \tag{5.47}$$

The proof is decomposed into two steps. First, we show that for all values of the p_j abiding by the constraints and for all weights $u_2, \dots, u_K$ such that $u_j \geq 0$ for all j and $u_1 + \ldots + u_K = 1$, there exists $k^\star \in \{2, \dots, K\}$ such that there exists an increasing sequence of budgets $(T_n)_{n \geq 1}$ with

$$\liminf_{n \to +\infty} \frac{1}{T_n} \log \mathbb{P}_{\underline{\nu}^{(k^\star)}}\big(\hat{a}_{T_n} \neq k^\star\big) \geq -9\, u_{k^\star} \big(\Delta_{k^\star}^{\mathsf{base}}\big)^2\,. \tag{5.48}$$

Then, we set specific values of the u_j and p_j to get

$$\forall k \in \{2, \dots, K\}, \qquad u_k \big(\Delta_k^{\mathsf{base}}\big)^2 \leq \frac{10}{3 \log K}\, H_\Sigma\big(\underline{\nu}^{(k)}\big)^{-1}\,. \tag{5.49}$$

Proposition 5.24 follows by combining (5.48) and (5.49).

Part 1: Proof of (5.48). For all $T \geq 1$,

$$\sum_{k=2}^{K} \frac{\mathbb{E}_{\underline{\nu}^{\mathsf{base}}}\big[N_k(T)\big]}{T} \leq 1 = \sum_{k=2}^{K} u_k\,;$$

therefore, for all $T \geq 1$, there exists $k_T \in \{2, \dots, K\}$ such that $\mathbb{E}_{\underline{\nu}^{\mathsf{base}}}\big[N_{k_T}(T)\big]/T \leq u_{k_T}$. By the pigeonhole principle, there exists $k^\star \in \{2, \dots, K\}$ and an (infinite) increasing sequence $(T_n)_{n \geq 1}$ of integers such that $k_{T_n} = k^\star$ for all $n \geq 1$. In particular,

$$\limsup_{n \to +\infty} \frac{\mathbb{E}_{\underline{\nu}^{\mathsf{base}}}\big[N_{k^\star}(T_n)\big]}{T_n} \leq u_{k^\star}\,.$$

Since $\underline{\nu}^{\mathsf{base}}$ and $\underline{\nu}^{(k^\star)}$ only differ at arm $k^\star$, an application of Lemma 5.7 along subsequences (see the initial comments below the lemma) guarantees that

$$\begin{aligned} \liminf_{n \to +\infty} \frac{1}{T_n} \log \mathbb{P}_{\underline{\nu}^{(k^\star)}}\big(\hat{a}_{T_n} \neq k^\star\big) &\geq -\left(\limsup_{n \to +\infty} \frac{\mathbb{E}_{\underline{\nu}^{\mathsf{base}}}\big[N_{k^\star}(T_n)\big]}{T_n}\right) \mathrm{KL}\big(\mathrm{Ber}(1 - p_{k^\star}), \mathrm{Ber}(p_{k^\star})\big) \\ &\geq -u_{k^\star} \times 9\,(1/2 - p_{k^\star})^2 = -9\, u_{k^\star}\, \big(\Delta_{k^\star}^{\mathsf{base}}\big)^2\,, \end{aligned}$$

where, in the last inequality, we used that for all $x \in [1/4,\, 1/2)$,

$$\mathrm{KL}(\mathrm{Ber}(1-x), \mathrm{Ber}(x)) = (1-x)\log\frac{1-x}{x} + x\log\frac{x}{1-x} \leq 9\left(\frac{1}{2}-x\right)^2 .$$

Part 2: Proof of (5.49). We set, for $j \in \{2,\dots,K\}$,

$$u_j = \frac{U}{\left(\Delta_j^{\mathsf{base}}\right)^2 H_\Sigma(\underline{\nu}^{(j)})}\,, \qquad \text{where} \qquad U = \left(\sum_{k=2}^{K} \frac{1}{\left(\Delta_k^{\mathsf{base}}\right)^2 H_\Sigma(\underline{\nu}^{(k)})}\right)^{-1} .$$

Then, $u_k\left(\Delta_k^{\mathsf{base}}\right)^2 = H_\Sigma(\underline{\nu}^{(k)})^{-1} U$ for all $k \in \{2,\dots,K\}$. To get the desired result, it suffices to guarantee that $U \leq 10/(3\log K)$. To do so, we consider the same values as in Carpentier and Locatelli (2016) for the p_j, i.e., we set, for $j \in \{2,\dots,K\}$,

$$p_j = \frac{1}{2} - \frac{j}{4K} \qquad \text{or, equivalently,} \qquad \Delta_j^{\mathsf{base}} = \frac{j}{4K}\,.$$

We show first that $\left(\Delta_k^{\mathsf{base}}\right)^2 H_\Sigma(\nu^{(k)}) \leq 2k$, for all $k \in \{2,\dots,K\}$. Indeed, by (5.47) and by lower bounding $\Delta_k^{\mathsf{base}} + \Delta_j^{\mathsf{base}}$ either by Δ_k^{base} or Δ_j^{base}, we get

$$\begin{aligned}
\left(\Delta_k^{\mathsf{base}}\right)^2 H_\Sigma(\underline{\nu}^{(k)}) &= \sum_{j<k} \frac{\left(\Delta_k^{\mathsf{base}}\right)^2}{\left(\Delta_k^{\mathsf{base}} + \Delta_j^{\mathsf{base}}\right)^2} + \sum_{j>k} \frac{\left(\Delta_k^{\mathsf{base}}\right)^2}{\left(\Delta_k^{\mathsf{base}} + \Delta_j^{\mathsf{base}}\right)^2} \\
&\leq k - 1 + \sum_{j>k} \frac{\left(\Delta_k^{\mathsf{base}}\right)^2}{\left(\Delta_j^{\mathsf{base}}\right)^2} = k - 1 + \sum_{j>k} \frac{k^2}{j^2} \leq k - 1 + k^2 \int_k^K \frac{1}{v^2}\mathrm{d}v \leq 2k\,.
\end{aligned}$$

Finally,

$$U \leq \left(\sum_{k=2}^{K} \frac{1}{2k}\right)^{-1} \leq \left(\int_2^{K+1} \frac{1}{2v}\mathrm{d}v\right)^{-1} = 2\left(\log(K+1) - \log 2\right)^{-1} \leq \frac{10}{3\log K}\,,$$

where the final inequality holds since $K \geq 3$. □

5.6.2. The Bretagnolle-Huber Technique by Kaufmann et al. (2016)

Kaufmann et al. (2016, Section 5.2) provide an interesting series of results relying on the so-called Bretagnolle-Huber inequality recalled below in (5.51); we state one of their lower bounds in Corollary 5.26. But as we argue in this section, the methodology followed seems extremely specific to the case of parametric models where Kullback-Leibler divergences could be controlled (lower bounded and upper bounded) in terms of gaps, like the model $\mathcal{D}_{\sigma^2}$ of Gaussian distributions with a fixed variance $\sigma^2 > 0$. In particular, we state in Proposition 5.25 what would be the straightforward extension to non-parametric models of the Gaussian results of (Kaufmann et al., 2016, Section 5.2), and we immediately discuss after this statement why this extension lacks interpretability and interest. Proposition 5.25 considers any sequence of strategies (not necessarily consistent) and provides an asymptotic bound; however, it does not directly control the target probability of error $\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))$, but a larger quantity. A proof of Proposition 5.25 is provided at the end of this section.

Proposition 5.25. *Fix $K \geq 2$, a model $\mathcal{D}$, and any sequence of strategies. Let $\underline{\nu}$ be a bandit problem in $\mathcal{D}$ with a unique optimal arm. Consider, for each $k \neq a^\star(\underline{\nu})$, a distribution $\zeta_k \in \mathcal{D}$ such that $\mathrm{E}(\zeta_k) > \mu^\star$. For $k \neq a^\star(\underline{\nu})$, denote by $\underline{\nu}^{(k)}$ the bandit problem obtained from $\underline{\nu}$ by changing the distribution of arm k into ζ_k. For all $T \geq 1$,*

$$\frac{1}{T}\log\max\left\{\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})),\ \max_{k\neq a^\star(\underline{\nu})}\mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T \neq k)\right\} \geq -\left(\sum_{a\neq a^\star(\underline{\nu})}\frac{1}{\mathrm{KL}(\nu_a,\zeta_a)}\right)^{-1} - \frac{\log 4}{T}\,.$$

Lack of interpretability of the bound for general models. To derive an interesting and interpretable bound from this result, one needs to choose carefully the distributions ζ_k. There is a tradeoff between obtaining a large lower bound by choosing ζ_k as close as possible to ν_k in terms of Kullback-Leibler divergences, and controlling the maximum of the misidentification probabilities: when ζ_k gets closer to ν_k while abiding by the constraint $\mathrm{E}(\zeta_k) > \mu^\star$, the probability $\mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T \neq k)$ becomes larger, and should even intuitively converge to $1/2$. In any case, the target error $\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))$ should get dominated by $\mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T \neq k)$ and the obtained bound is likely to be uninformative on the target error, due to the maximum on the left-hand side. This tradeoff seems to be unsolvable in general unless there exist some specific properties for the Kullback-Leibler divergence of the model, as we illustrate below for a Gaussian model, which was the setting considered by Kaufmann et al. (2016, Section 5.2).

Another intuitive issue with the bound of Proposition 5.25 is that it involves Kullback-Leibler divergences with arguments in reverse order compared to the lower bounds presented in Section 5.4. Indeed, taking the supremum of the lower bound over distributions ζ_k such that $\mathrm{E}(\zeta_k) > \mu^\star$ would lead to a complexity in terms of the $\mathcal{K}^{>}_{\inf}(\nu_k, \mu^\star)$, where

$$\mathcal{K}^{>}_{\inf}(\nu, x) \stackrel{\text{def}}{=} \inf\{\mathrm{KL}(\nu,\zeta) : \zeta \in \mathcal{D} \ \text{ s.t. } \ \mathrm{E}(\zeta) > x\}\,,$$

rather than in terms of the $\mathcal{L}^{>}_{\inf}(\mu^\star, \nu_k)$. Our intuition, given all bounds presented in this chapter, is that the $\mathcal{K}^{>}_{\inf}(\nu_k, \mu^\star)$ would not form the correct notion of complexity for the fixed-budget best-arm identification.

How Kaufmann et al. (2016, Section 5.2) could exploit Proposition 5.25 in the Gaussian case. Yet, in the case of the model $\mathcal{D}_{\sigma^2}$ of Gaussian distributions with a fixed variance $\sigma^2 > 0$, for which KL is symmetric, Proposition 5.25 admits an interesting corollary, corresponding[4] to Theorem 16 of Kaufmann et al. (2016, Section 5.2). The corollary actually relies on a strong property of KL in this model: not only is it symmetric, but it only depends on the expectation gaps between its arguments. Namely, for all pairs $\mathcal{N}(\mu,\sigma^2)$ and $\mathcal{N}(\mu',\sigma^2)$ of distributions in $\mathcal{D}_{\sigma^2}$, for all $\Delta \in \mathbb{R}$,

$$\mathrm{KL}(\mathcal{N}(\mu,\sigma^2), \mathcal{N}(\mu',\sigma^2)) = \frac{(\mu-\mu')^2}{2\sigma^2} = \mathrm{KL}(\mathcal{N}(\mu+\Delta,\sigma^2), \mathcal{N}(\mu'+\Delta,\sigma^2))\,. \tag{5.50}$$

We introduce the following short-hand notation:

$$H_\Sigma(\underline{\nu}) \stackrel{\text{def}}{=} \sum_{a\neq a^\star(\underline{\nu})}\frac{2\sigma^2}{\Delta_a^2}\,.$$

[4] The maximum of the left-hand side of Corollary 5.26 is present, but somewhat discrete, in the Theorem 16 of Kaufmann et al. (2016, Section 5.2): it corresponds to the "There exists an alternative bandit problem" part of the statement of the latter.

Corollary 5.26. *For all sequences of strategies and for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{\sigma^2}$ with a unique optimal arm, there exists a set of alternative bandit instances $(\underline{\nu}^{(k)})_{k\neq a^\star(\underline{\nu})}$ in $\mathcal{D}_{\sigma^2}$, where each $\underline{\nu}^{(k)}$ admits k as a best arm and satisfies $H_\Sigma(\underline{\nu}^{(k)}) \leq H_\Sigma(\underline{\nu})$, and for which*

$$\frac{1}{T}\log\max\Big\{\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})),\ \max_{k\neq a^\star(\underline{\nu})}\mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T\neq k)\Big\} \geq -4\,H_\Sigma(\underline{\nu})^{-1} - \frac{\log 4}{T}\,.$$

The proof provided below is highly specific to the Gaussian model and exploits the gap-based rewriting (5.50) of the Kullback-Leibler divergence. The calculations led would only extend to models for which such gap-based rewritings of (upper and lower bounds on) the Kullback-Leibler divergence would be available.

To compare the result of Corollary 5.26 with the bound (5.8) stemming from Audibert et al. (2010), note that

$$H_\Sigma(\underline{\nu})^{-1} \geq \frac{2}{\sigma^2}\min_{2\leq k\leq K}\frac{\Delta_{(k)}^2}{k}\,.$$

Proof. We apply Proposition 5.25 with the distributions $\zeta_k = \mathcal{N}(\mu^\star+\Delta_k,\sigma^2)$, for $k\neq a^\star(\underline{\nu})$. On the one hand, the bound of Proposition 5.25 involves

$$\sum_{a\neq a^\star(\underline{\nu})}\frac{1}{\mathrm{KL}(\nu_a,\zeta_a)} = \sum_{a\neq a^\star(\underline{\nu})}\frac{2\sigma^2}{\big(\underbrace{\mathrm{E}(\nu_a)}_{\mu^\star-\Delta_a}-\underbrace{\mathrm{E}(\zeta_a)}_{\mu^\star+\Delta_a}\big)^2} = \frac{H_\Sigma(\underline{\nu})}{4}\,.$$

On the other hand, for $k\neq a^\star(\underline{\nu})$, as the best arm of $\underline{\nu}^{(k)}$ is k, with associated expectation $\mu^\star+\Delta_k$,

$$\begin{aligned}H_\Sigma(\underline{\nu}^{(k)}) = \sum_{a\neq k}\frac{2\sigma^2}{(\mu^\star+\Delta_k-\mu_a)^2} &= \frac{2\sigma^2}{\Delta_k^2} + \sum_{a\notin\{k,a^\star(\underline{\nu})\}}\frac{2\sigma^2}{(\mu^\star+\Delta_k-\mu_a)^2}\\ &\leq \frac{2\sigma^2}{\Delta_k^2} + \sum_{a\notin\{k,a^\star(\underline{\nu})\}}\frac{2\sigma^2}{(\mu^\star-\mu_a)^2} = \sum_{a\neq a^\star(\underline{\nu})}\frac{2\sigma^2}{\Delta_a^2} = H_\Sigma(\underline{\nu})\,.\end{aligned}$$

These two observations conclude the proof of Corollary 5.26. □

Proof of Proposition 5.25. We conclude this section with a proof of Proposition 5.25. It relies on the Bretagnolle-Huber inequality (Bretagnolle and Huber, 1979), which states that, for all $p,q\in[0,1]$,

$$p+1-q \geq \frac{1}{2}\exp\Big(-\mathrm{KL}(\mathrm{Ber}(p),\mathrm{Ber}(q))\Big)\,. \tag{5.51}$$

Proof. We fix distributions ζ_k abiding by the conditions of the proposition and also fix $T\geq 1$. We will prove below that, for all convex weights $(u_b)_{b\neq a^\star(\underline{\nu})}$, i.e., non-negative weights summing up to 1,

$$\frac{1}{T}\log\max\Big\{\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})),\ \max_{k\neq a^\star(\underline{\nu})}\mathbb{P}_{\underline{\nu}^{(k)}}(\hat{a}_T\neq k)\Big\} \geq -\max_{b\neq a^\star(\underline{\nu})}\{u_b\,\mathrm{KL}(\nu_b,\zeta_b)\} - \frac{\log 4}{T}\,, \tag{5.52}$$

from which Proposition 5.25 follows, by optimizing the obtained lower bound, i.e., by taking

$$u_b = \left(\sum_{a\neq a^\star(\underline{\nu})}\frac{1}{\mathrm{KL}(\nu_a,\zeta_a)}\right)^{-1}\times\frac{1}{\mathrm{KL}(\nu_b,\zeta_b)}\,.$$

We now fix convex weights $(u_b)_{b\neq a^\star(\underline{\nu})}$ and prove (5.52). As $b \neq a^\star(\underline{\nu})$ and b is the unique optimal arm of $\underline{\nu}^{(b)}$, for the first inequality, and by the Bretagnolle-Huber inequality (5.51), for the second inequality,

$$\begin{aligned}\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) + \mathbb{P}_{\underline{\nu}^{(b)}}(\hat{a}_T \neq b) &\geq \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})) + \mathbb{P}_{\underline{\nu}^{(b)}}(\hat{a}_T = a^\star(\underline{\nu})) \\ &\geq \frac{1}{2}\exp\Big(-\operatorname{KL}(\operatorname{Ber}(p_T), \operatorname{Ber}(q_T))\Big)\,,\end{aligned}$$

where $p_T \overset{\text{def}}{=} \mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu}))$ and $q_T \overset{\text{def}}{=} \mathbb{P}_{\underline{\nu}^{(b)}}(\hat{a}_T \neq a^\star(\underline{\nu}))$. Inequality (2.12) reads, in the present case, as $\underline{\nu}$ and $\underline{\nu}^{(b)}$ only differ at arm b,

$$\operatorname{KL}(\operatorname{Ber}(p_T), \operatorname{Ber}(q_T)) \leq \mathbb{E}_{\underline{\nu}}[N_b(T)]\ \operatorname{KL}(\nu_b, \zeta_b)\,.$$

Using $\max\{u, v\} \geq (u+v)/2$ after collecting all bounds obtained so far yields

$$\max\Big\{\mathbb{P}_{\underline{\nu}}(\hat{a}_T \neq a^\star(\underline{\nu})), \mathbb{P}_{\underline{\nu}^{(b)}}(\hat{a}_T \neq b)\Big\} \geq \frac{1}{4}\exp\Big(-\mathbb{E}_{\underline{\nu}}[N_b(T)]\ \operatorname{KL}(\nu_b, \zeta_b)\Big)\,.$$

We take the maxima over $b \neq a^\star(\underline{\nu})$ in both sides, apply logarithms, and conclude the proof of (5.52) by showing that

$$\min_{b\neq a^\star(\underline{\nu})}\Big\{\mathbb{E}_{\underline{\nu}}[N_b(T)]\ \operatorname{KL}(\nu_b, \zeta_b)\Big\} \leq \max_{b\neq a^\star(\underline{\nu})}\{u_b\ \operatorname{KL}(\nu_b, \zeta_b)\}\,. \tag{5.53}$$

Indeed,

$$\sum_{b\neq a^\star(\underline{\nu})} \frac{\mathbb{E}_{\underline{\nu}}[N_b(T)]}{T} \leq 1 = \sum_{b\neq a^\star(\underline{\nu})} u_b\,,$$

so that there exists $b^\star \neq a^\star(\underline{\nu})$ such that $\mathbb{E}_{\underline{\nu}}[N_{b^\star}(T)]/T \leq u_{b^\star}$. We then have

$$\min_{b\neq a^\star(\underline{\nu})}\Big\{\mathbb{E}_{\underline{\nu}}[N_b(T)]\ \operatorname{KL}(\nu_b, \zeta_b)\Big\} \leq u_{b^\star}\ \operatorname{KL}(\nu_{b^\star}, \zeta_{b^\star}) \leq \max_{b\neq a^\star(\underline{\nu})}\{u_b\ \operatorname{KL}(\nu_b, \zeta_b)\}\,,$$

as desired in (5.53). □

5.7. Conclusion

In this chapter, we explored the challenging fixed-budget setting of best-arm identification. We introduced new tools, including the information-theoretic quantities $\mathcal{L}^{<}_{\inf}$, $\mathcal{L}^{\leq}_{\inf}$, $\mathcal{L}^{>}_{\inf}$ and $\mathcal{L}^{\geq}_{\inf}$, in order to generalize the gap-based bounds of the literature to more general models. Our new general analysis of the `Successive-Rejects` strategy indicates that the complexity of the fixed-budget setting might be measured by the $\mathcal{L}$ quantity defined as

$$\mathcal{L}(\nu', \nu) = \inf_{x\in[\mathrm{E}(\nu'),\mathrm{E}(\nu)]}\Big\{\mathcal{L}^{\geq}_{\inf}(x, \nu') + \mathcal{L}^{\leq}_{\inf}(x, \nu)\Big\}\,.$$

We then stated several lower bounds depending on various assumptions. Those bounds do not express in terms of the $\mathcal{L}$ quantity —although the bound of Theorem 5.13 invokes quantities that are getting closer to $\mathcal{L}$—, but we hope that future works could bring new ideas to improve those bounds.

Existence of a complexity As previously explained, there is still a gap of at least a factor $\overline{\log} K$ between the lower and upper bounds presented in this chapter. In fact, we do not know if we can fill this gap and hence prove the existence of a complexity like in the fixed-confidence setting. Recent works (see Komiyama et al., 2022; Degenne, 2023) have been focusing on proving that such optimal complexity does not exist for large enough models (see Section 2.3.5 for discussions).

CHAPTER 6

Asymptotically Optimal Adaptive Top-Two Algorithms in the Fixed-Confidence Setting

This chapter studies adaptive `top-two` algorithms for a general exponential model in the fixed-confidence setting. It gathers preliminary works that have not been submitted to date. Mainly two conjectures are stated, supported by numerical experiments and elements of proof. Firstly, we obtain new procedures to compute optimal weight vectors that appear to be numerically efficient. Secondly, we give a proof structure to generalize the asymptotic optimality of adaptive `top-two` algorithms obtained for a Gaussian model by You et al. (2023).

Contents

6.1. Introduction

We consider here the problem of best-arm identification with a fixed-confidence (see Section 2.2 for motivations about this setting and for more details concerning the following introduction). Given some confidence parameter $\delta \in (0,1)$, our aim is to find a strategy that minimizes the expected number of samplings $\mathbb{E}_{\underline{\nu}}[\tau_\delta]$ among δ–correct strategies, i.e., strategies such that for all bandit problems $\underline{\nu}$,

$$\mathbb{P}_{\underline{\nu}}(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^\star(\underline{\nu})) \leq \delta\,.$$

When considering an exponential model $\mathcal{D}_{\exp}$ (see the reminder on page 43), we recall that bandit problems $\underline{\nu}$ are characterized by their mean vector $\underline{\mu}$, and a notion of asymptotic optimality (when δ goes to 0) was introduced for the identification problem by Garivier and Kaufmann (2016). They first proved that all δ–correct strategies satisfy, for all bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$,

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log\frac{1}{\delta}} \geq T(\underline{\mu})\,, \tag{6.1}$$

where $T(\underline{\mu})$ denotes the *characteristic time* of $\underline{\mu}$, defined as

$$T(\underline{\mu})^{-1} \stackrel{\text{def}}{=} \sup_{\underline{v}\in\Sigma_K} \inf_{\underline{\lambda}\in \mathrm{Alt}(\underline{\mu})} \sum_{a\in[K]} v_a\, d(\mu_a,\lambda_a)\,, \tag{6.2}$$

where d denotes the mean-parameterized Kullback-Leibler divergence of the model $\mathcal{D}_{\exp}$,

$$\Sigma_K = \Big\{\underline{v}\in[0,1]^K \,:\, v_1+\cdots+v_K = 1\Big\} \qquad \text{and} \qquad \mathrm{Alt}(\underline{\mu}) = \Big\{\underline{\lambda} \text{ in } \mathcal{D}_{\exp} \,:\, a^\star(\underline{\lambda}) \neq a^\star(\underline{\mu})\Big\}\,.$$

Additionally, they designed `Track-and-Stop` , the first *asymptotically optimal strategy*, for which the asymptotic upper bound matches the above lower bound:

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\exp}, \qquad \limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log\frac{1}{\delta}} \leq T(\underline{\mu})\,.$$

The existence of such strategies ensures the asymptotic tightness of lower bound (6.1).

Computing the solution $\underline{w}(\underline{\mu})$. The information-theoretic analysis of Garivier and Kaufmann (2016) also highlights the nature of optimal sampling strategies: whatever the value of the risk δ, they should sample the arms with frequencies proportional to $\underline{v} = \underline{w}(\underline{\mu})$, the (unique and well-defined) maximizer of optimization problem (6.2), called the *optimal weight vector*. To achieve this, the `Track-and-Stop` sampling rule estimates these proportions by computing, at each time step t, the optimal weight vector $\underline{w}(\underline{\hat{\mu}}(t))$ of the current empirical mean $\underline{\hat{\mu}}(t)$. Indeed, they proved that, with the knowledge of $\underline{\mu}$, solving the optimization problem (6.2) reduces to determining the root of a one-variable increasing function. By applying a bisection method, one may then compute $\underline{w}(\underline{\mu})$ with arbitrary precision. Yet, it can be interesting to obtain more efficient methods to compute this optimal weight vector, as we did for a Gaussian model in Chapter 3.

The shortcomings of `Track-and-Stop`. Computing the solution of optimization problem (6.2) at each time step is not very efficient in terms of computational cost. This is one of the shortcomings of `Track-and-Stop` (see Section 2.2.5), we may also cite, among others, the requirement of forced exploration to ensure that all arms are pulled sufficiently. Some improvements were proposed: for example, Ménard (2019) and Wang et al. (2021) proved that it is not necessary to solve the optimization problem at every time step. Instead, they perform a single gradient step in every round, which enables asymptotic optimality of computationally more efficient algorithms. Another direction of simplification is the study of `top-two` algorithms (see Russo, 2016) that we present in depth below.

The importance of sampling rules. The performance of a strategy highly depends on both its sampling and stopping rules. Yet, the analysis of Garivier and Kaufmann (2016) proved (see Theorem 2.9) that the δ–correctness of a strategy can be obtained, whatever the sampling rule is, using the `Global-Likelihood-Ratio` stopping rule with a carefully designed threshold:

$$\beta(t, \delta) \stackrel{\text{def}}{=} \log \frac{Rt^\alpha}{\delta}\,, \tag{6.3}$$

where $\alpha > 1$ and R is a constant depending on α and K. As it seems that the stopping rule cannot be significantly improved, the performance of a strategy should mostly be determined by its sampling rule.

`top-two` **algorithms.** `top-two` strategies are promising sets of strategies that come with simple sampling rules: at time step t, the algorithm chooses the next arm to sample A_t between two arms, namely a *leader* L_t and a *challenger* C_t. We recall the general structure of the sampling rule in Algorithm 17. An example of a natural leader is to choose the arm with the current best empirical mean (Algorithm 18), while the challenger might be the arm minimizing its *transportation cost*, defined in Equation (6.6), with the leader, or a penalized version of that cost which encourages exploration (Algorithm 19). See also Section 2.2.7 (and references therein) for additional `leader` and `challenger` procedures.

Algorithm 17: `top-two` sampling rule at time step $t > K$

Input: history of observations I_{t-1}
`leader`, `challenger`, `sample-arm` procedures
Output: next arm to observe A_t

$L_t \leftarrow \texttt{leader}(I_{t-1})$
$C_t \leftarrow \texttt{challenger}(I_{t-1}, L_t)$
$A_t \leftarrow \texttt{sample-arm}(I_{t-1}, L_t, C_t)$

Algorithm 18: `Empirical-Best` leader

Input: history of observations I_{t-1}
Output: leader L_t

Choose $L_t \in \operatorname*{argmax}_{a \in [K]} \hat{\mu}_a(t-1)$

Algorithm 19: `challenger` procedures for `top-two` algorithms

Input: history of observations I_{t-1}
leader L_t
Output: challenger C_t

/* `Transportation-Cost (TC)` challenger */
Choose $C_t \in \operatorname*{argmin}_{a \neq L_t} \mathrm{TC}_{a \to L_t}\left(\underline{\hat{\mu}}(t-1), \frac{\underline{N}(t-1)}{t-1}\right)$
// $\mathrm{TC}_{a\to b}(\mu, \underline{\nu})$ is defined in (6.6)
/* `Transportation-Cost-Penalized (TCP)` challenger */
Choose $C_t \in \operatorname*{argmin}_{a \neq L_t} \mathrm{TC}_{a \to L_t}\left(\underline{\hat{\mu}}(t-1), \frac{\underline{N}(t-1)}{t-1}\right) + \log N_a(t-1)$

The asymptotic optimality of `top-two` algorithms has been discussed, but only (until very recently) for non-adaptive strategies, for which the sampling rule chooses the leader with a fixed probability parameter $\beta \in (0,1)$. As a good `leader` might satisfy $L_t = a^\star$ except for a sub-linear number of time steps, such non-adaptive strategies will pull the best arm a fraction β of the time, which can be far from the optimal frequency $w_{a^\star}(\underline{\mu})$. As a consequence, those strategies satisfy the following lower bound (see Russo, 2016):

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\exp}, \qquad \liminf_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \geq T_\beta(\underline{\mu}) \stackrel{\text{def}}{=} \left(\sup_{\substack{\underline{v} \in \Sigma_K \\ v_{a^\star} = \beta}} \inf_{\lambda \in \mathrm{Alt}(\underline{\mu})} \sum_{a \in [K]} v_a \, d(\mu_a, \lambda_a) \right)^{-1}.$$

It has been proved (see, e.g., Jourdan et al., 2022) that good choices of leaders and challengers (like those proposed in Algorithms 18 and 19) lead to β–asymptotically optimal strategies, that is, strategies such that, for all generic (hence with distinct means) bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T_\beta(\underline{\mu}) \, .$$

Adaptive `top-two` algorithms. Non-adaptive algorithms are hence asymptotically optimal only if $\beta = w_{a^\star}(\underline{\mu})$, which is quite frustrating even if it can be proved (see Russo, 2016) that $T_{\frac{1}{2}}(\underline{\mu}) \leq 2T(\underline{\mu})$, i.e., that the choice of $\beta = \frac{1}{2}$ leads to a loss in guarantee of a multiplicative factor 2 only. To tackle this problem, we might consider adaptive algorithms, for which the sampled arm between the leader and the challenger is not chosen according to an external parameter, but thanks to an adaptive parameter. Based on this idea, You et al. (2023) proved that `Top-Two-Thompson-Sampling` with an adaptive sampling rule (a `top-two` strategy with a Bayesian choice of the leader and the challenger, see Section 2.2.7) is asymptotically optimal, but only for a Gaussian model with common variance $\sigma^2 > 0$.

Outline and contributions. The question of extending the asymptotic guarantees of adaptive `top-two` algorithms to more general models is still open. The objective of this chapter is to present preliminary work in that direction. We focus on exponential models, which form the class of models for which the fixed-confidence best-arm identification problem is best understood (see Garivier and Kaufmann, 2016).

To begin with, in Section 6.2, we prove that for a general exponential model, the solution $\underline{w}(\underline{\mu})$ of optimization problem (6.2) can be seen as the unique fixed point of some transformation. We use this property to present a new procedure for computing the optimal weight vector $\underline{w}(\underline{\mu})$ that turns out to be empirically efficient. Additionally, an interesting interpretation of the transformation naturally leads to the introduction of a new `challenger` rule together with an adaptive sampling rule for `top-two` algorithms. Then, in Section 6.3, we explore how to generalize the Gaussian analysis of You et al. (2023) to a general exponential model $\mathcal{D}_{\exp}$. As a by-product of the analysis of Section 6.3, we give a new proof of the asymptotic optimality of `Track-and-Stop` in Section 6.4.

Notation. For a given strategy facing a bandit problem $\underline{\mu}$, let $N_a(t)$ and $\hat{\mu}_a(t)$ denote the number of pulls and the empirical mean[1] of arm a at step t:

$$N_a(t) \stackrel{\text{def}}{=} \sum_{s \in [t]} \mathbb{I}\{A_s = a\} \qquad \text{and} \qquad \hat{\mu}_a(t) \stackrel{\text{def}}{=} \frac{1}{N_a(t)} \sum_{s \in [t]} Y_s \, \mathbb{I}\{A_s = a\} \, .$$

[1] As strategies initially observe each arm once, $\hat{\mu}_a(t)$ is well-defined for $t \geq K$.

In the rest of this chapter, we fix some exponential model $\mathcal{D}_{\exp}$ and rely on the associated notation defined on page 43. In a nutshell, distributions ν_θ of the model are parameterized by elements θ of the natural space parameter Θ, or equivalently by their means $\mu = \mathrm{E}(\nu_\theta)$, which belong to a convex open interval $\mathcal{M} = (\mu_-, \mu_+)$. The log–partition function, denoted by b, is twice differentiable, such that $\mathrm{E}(\nu_\theta) = b'(\theta)$ and b' is invertible, and the mean-parameterized Kullback-Leibler divergence function of the model is defined, for all $\theta, \theta' \in \Theta$, by

$$d(\mathrm{E}(\nu_\theta), \mathrm{E}(\nu_{\theta'})) \stackrel{\text{def}}{=} \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = (\theta - \theta')b'(\theta) - b(\theta) + b(\theta') \,. \tag{6.4}$$

6.2. A Fixed Point Property

Let, until the end of Section 6.2.2, $\underline{\mu}$ be a fixed bandit instance in $\mathcal{D}_{\exp}$ with a unique best arm $a^\star$, and $\underline{w} = \underline{w}(\underline{\mu})$ be its optimal weight vector. We focus on optimization problem (6.2) and recall its formulation in terms of transportation costs (see Section 2.2.4):

$$T(\underline{\mu})^{-1} = \sup_{\underline{v} \in \Sigma_K} \min_{a \neq a^\star} \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v}) \,, \tag{6.5}$$

where, for $\underline{v} \in \Sigma_K$ and $a \neq a^\star$,

$$\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v}) = v_{a^\star}\, d(\mu^\star, \overline{\mu}_{a^\star, a, \underline{v}}) + v_a\, d(\mu_a, \overline{\mu}_{a^\star, a, \underline{v}}) \,, \qquad \text{where } \overline{\mu}_{a^\star, a, \underline{v}} \stackrel{\text{def}}{=} \frac{v_{a^\star}\mu^\star + v_a \mu_a}{v_{a^\star} + v_a} \,. \tag{6.6}$$

Garivier and Kaufmann (2016) proved that the optimal weight vector $\underline{w}$ solving (6.5) is characterized by the following sufficient and necessary conditions, where $\mathrm{int}(A)$ is the interior of a set A.

Proposition 6.1. *Let $\underline{\mu}$ be a fixed bandit instance in $\mathcal{D}_{\exp}$ with a unique best arm $a^\star$. Then for all $\underline{v} \in \mathrm{int}(\Sigma_K)$, the following conditions are equivalent:*

(i) $\underline{v} = \underline{w}(\underline{\mu})$,

(ii) for all sub-optimal arms $a \neq a^\star$,

$$\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w}) = \left(\sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{a^\star, b, \underline{w}})} \right)^{-1} = T(\underline{\mu})^{-1} \,, \tag{6.7}$$

(iii) all transportation costs $\left(\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w})\right)_{a \neq a^\star}$ *are equal and*

$$\sum_{b \neq a^\star} \frac{d(\mu^\star, \overline{\mu}_{a^\star, b, \underline{w}})}{d(\mu_b, \overline{\mu}_{a^\star, b, \underline{w}})} = 1 \,. \tag{6.8}$$

In Section 2.2.4, we explained the intuition why, for a given proportion $\beta \in (0,1)$, there exists a unique vector $\underline{w}^\beta = \underline{w}^\beta(\underline{\mu})$ such that $w^\beta_{a^\star} = \beta$ and which equalizes the transportation costs $\left(\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w}^\beta)\right)_{a \neq a^\star}$. We also explained that the optimal weight vector $\underline{w}$ belongs to the set of β–optimal weight vectors $\{\underline{w}^\beta : \beta \in (0,1)\}$, and that the corresponding value of β might be computed by a first-order condition. Conditions (6.7) and (6.8) are equivalent to this condition.

We use Proposition 6.1 to interpret $\underline{w}$ as the unique fixed point of some transformation in Section 6.2.1. We then analyze this transformation to design efficient procedures for the calculation of $\underline{w}$ in Section 6.2.2, and to introduce a new adaptive `top-two` sampling scheme in Section 6.2.3.

6.2.1. The Transformation

We recall that $\underline{\mu}$ is a fixed and known bandit instance in this section. At the optimum $\underline{w}$, thank to condition (6.7), all transportation costs are equal and we get

$$T(\underline{\mu}) = \sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{a^\star, b, \underline{w}})} \, .$$

Defining, for a weight vector $\underline{v} \in \mathrm{int}(\Sigma_K)$,

$$T_{\underline{v}} \stackrel{\mathrm{def}}{=} \sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{a^\star, b, \underline{v}})} \, , \tag{6.9}$$

the previous equation reads $T_{\underline{w}} = T(\underline{\mu})$. As $\underline{v} \mapsto T_{\underline{v}}$ is continuous, $T_{\underline{v}}$ can be interpreted as an approximation of $T(\underline{\mu})$, at least when $\underline{v}$ is around $\underline{w}$.

Consider some frequency vector $\underline{v} \in \Sigma_K$. If it is not optimal, one could try to correct the difference between costs $\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w})$ and $T_{\underline{v}}^{-1}$ for all sub-optimal arms a. As the cost $\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w})$ is homogeneous in $(v_{a^\star}, v_a)$, multiplying both the values of $v_{a^\star}$ and v_a by $T_{\underline{v}}^{-1} / \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})$ would transform the cost to the exact value $T_{\underline{v}}^{-1}$. Yet, note that if this multiplicative factor is not 1, changing the value of $(v_{a^\star}, v_a)$ for a fixed a will result in a vector out of Σ_K. Similarly, it is not possible to change simultaneously the value of the pair for all sub-optimal arm a, as $v_{a^\star}$ would be multiplied by different factors. However, we can see what will give the transformation which consists in updating each v_a by the corresponding multiplicative factor and then choosing $v_{a^\star}$ so as to get $\underline{v} \in \Sigma_K$. Formally, this defines a transformation $\underline{W} : \mathrm{int}(\Sigma_K) \to \mathrm{int}(\Sigma_K)$ by

$$\forall \underline{v} \in \mathrm{int}(\Sigma_K), \forall a \in [K], \qquad W_a(\underline{v}) \stackrel{\mathrm{def}}{=} \begin{cases} v_a \frac{T_{\underline{v}}^{-1}}{\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v})} & \text{if } a \neq a^\star \, , \\ 1 - \sum_{b \neq a^\star} v_b \frac{T_{\underline{v}}^{-1}}{\mathrm{TC}_{b \to a^\star}(\underline{\mu}, \underline{v})} & \text{if } a = a^\star \, . \end{cases} \tag{6.10}$$

It is easy to check that this transformation is well-defined. Indeed, for all $\underline{v} \in \mathrm{int}(\Sigma_K)$, we get that $W_{a^\star}(\underline{v}) > 0$ by using the definition of the costs (6.6):

$$\sum_{b \neq a^\star} \frac{v_b}{\mathrm{TC}_{b \to a^\star}(\underline{\mu}, \underline{v})} < \sum_{b \neq a^\star} \frac{v_b}{v_b \, d(\mu_b, \overline{\mu}_{b, a^\star, \underline{v}})} = \sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{b, a^\star, \underline{v}})} = T_{\underline{v}} \, .$$

By the necessary and sufficient conditions (6.7), the following fixed-point property holds.

Proposition 6.2. *Let $\underline{\mu}$ be a bandit instance in $\mathcal{D}_{\mathrm{exp}}$. Then $\underline{w}$ is the unique fixed point of $\underline{W}$.*

Proof. Using (6.7), one gets that, for all $a \neq a^\star$,

$$W_a(\underline{w}) = w_a \frac{T_{\underline{w}}^{-1}}{\mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{w})} = w_a \, ,$$

which ensures that $\underline{W}(\underline{w}) = \underline{w}$.

Reciprocally, if some vector $\underline{v} \in \mathrm{int}(\Sigma_K)$ is such that $\underline{W}(\underline{v}) = \underline{v}$, then

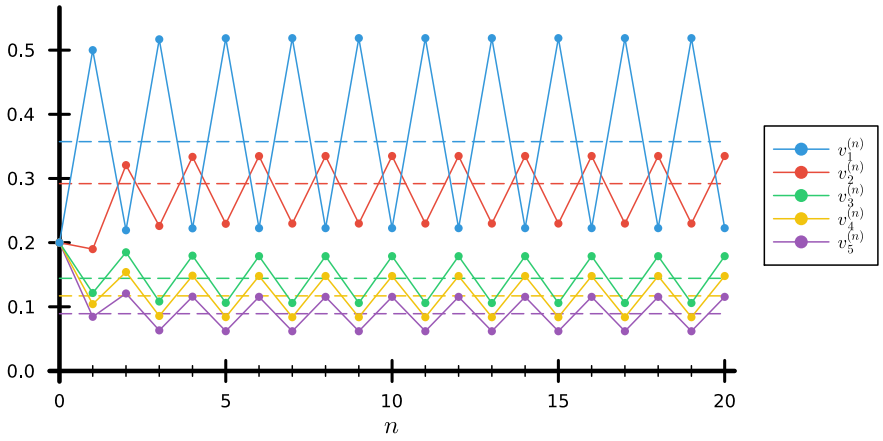
$$\forall a \neq a^\star, \qquad \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{v}) = T_{\underline{v}}^{-1} = \left(\sum_{b \neq a^\star} \frac{1}{d(\mu_b, \overline{\mu}_{a^\star, b, \underline{w}})} \right)^{-1} ,$$

hence $\underline{v}$ satisfies condition (6.7), which gives $\underline{v} = \underline{w}$. □

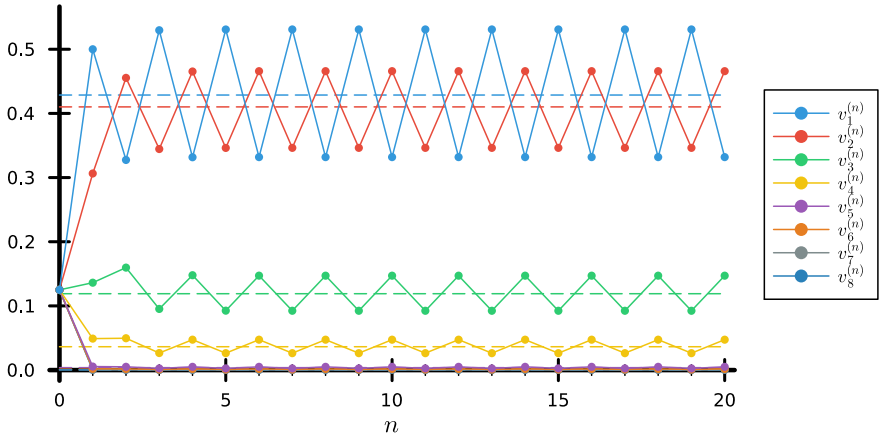
6.2.2. New Empirical `Optimal-Weights` Procedures

The computation of the optimal weight vector at each time step is the most costly part of strategies like `Track-and-Stop`. Reducing the complexity of a procedure `Optimal-Weights` which computes (an approximation of) those weights is then of high interest. We already proposed, independently, such an improvement for Gaussian variables in Chapter 3, and will now see how to use transformation $\underline{W}$ to define a procedure for an exponential model $\mathcal{D}_{\exp}$.

In view of the definition $\underline{W}$ and of its associated fixed point property for $\underline{w}$, it is natural to wonder whether the iterates $(\underline{v}^{(n)} \stackrel{\text{def}}{=} \underline{W}^n(\underline{v}^{(0)}))_{n\geq 0}$ of some initial vector $\underline{v}^{(0)}$ (e.g., the uniform vector) will converge to $\underline{w}$, by showing for instance that $\underline{W}$ is a contraction. If so, then we obtain a new `Optimal-Weights` procedure to approximate $\underline{w}$, which seems to be computationally efficient.



(a) $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$



(b) $\underline{\mu} = (0.95, 0.93, 0.92, 0.9, 0.8, 0.7, 0.5, 0.4)$

Figure 6.1: For the standard Gaussian model, coordinates of the iterated vectors $(\underline{v}^{(n)})_{n\geq 0}$ with $\underline{v}^{(0)}$ the uniform vector; the values of the corresponding coordinates of $\underline{w}$ are dashed.

$\underline{W}$ is not a contraction. Unfortunately, the convergence does not hold in all generality. Even in the simple model of standard Gaussian variables, the iterates may end up oscillating between two vectors and not converge to $\underline{w}$, see Figure 6.1. Let us highlights a few remarks concerning the numerical behavior that we observed whatever the Gaussian bandit instance. First, we observe that, when the

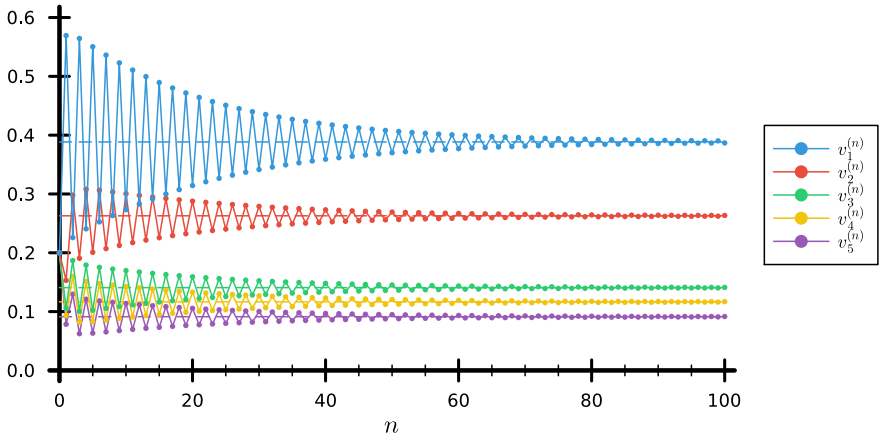


Figure 6.2: For the Bernoulli model and $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$, coordinates of the iterated vectors $(\underline{v}^{(n)})_{n\geq 0}$ with $\underline{v}^{(0)}$ the uniform vector.

weight of the optimal arm is over-estimated, those of sub-optimal arms are under-estimated, and vice versa. Then, the process quickly stabilizes into an oscillation between two vectors, which, strangely, do not satisfy any remarkable property, in particular:

- neither the optimal weight vector is the average of the two oscillations vectors, and, as a consequence, even the Cesàro sums of the iterates would not converge to $\underline{w}$,
- nor those two vectors equalize the transportation costs of sub-optimal arms. Nevertheless, they are numerically quite close to vectors satisfying this property, i.e., belonging to the set $\{\underline{w}^{\beta}(\underline{\mu}) : \beta \in (0,1)\}$ of β–optimal weight vectors. This explains why the weights of sub-optimal arms are approximately multiplied by the same value at each iteration (see Figure 6.1).

Remark. For Bernoulli instances, we observe in our experiments that the magnitude of the oscillations decreases slowly, and the iterates converge to the optimal weight vector, as can be seen in Figure 6.2. However, the convergence is not very fast and smooth.

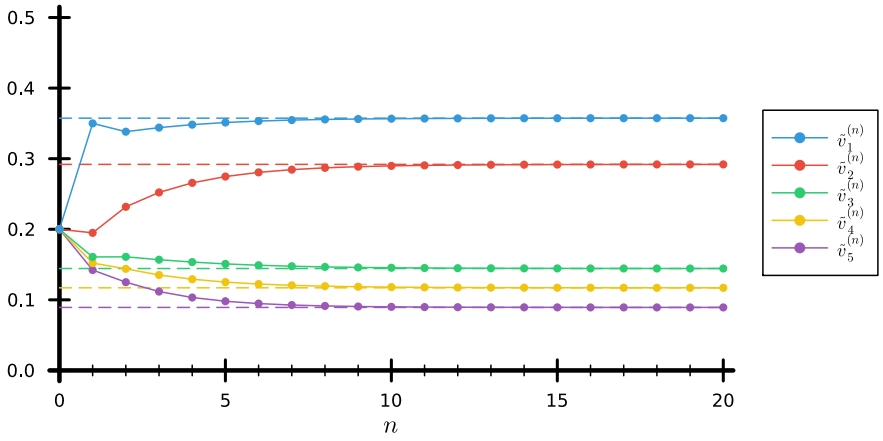
Soft update rule. In order to obtain convergence, one can consider less aggressive update rules. For instance, let us focus on the sequence defined by

$$\forall n \geq 0, \qquad \tilde{\underline{v}}^{(n+1)} \stackrel{\text{def}}{=} \frac{\tilde{\underline{v}}^{(n)} + \underline{W}(\tilde{\underline{v}}^{(n)})}{2}. \tag{6.11}$$

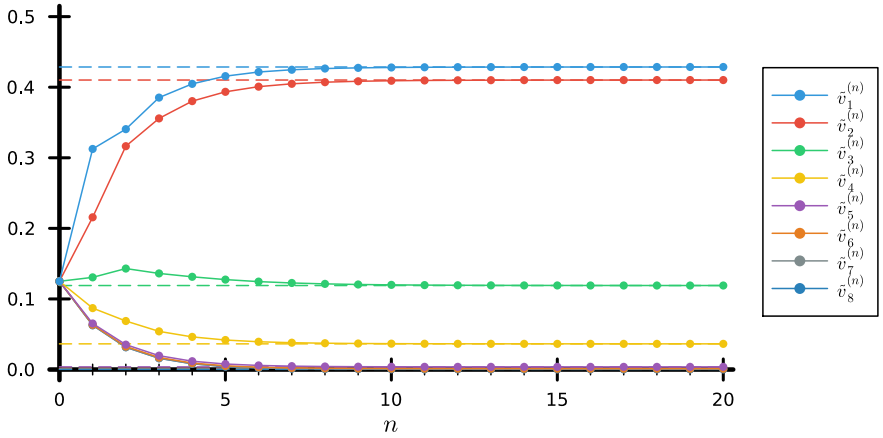
For this sequence, we observed on all our simulations that the convergence to $\underline{w}$ is quick and unambiguous whatever the model (see Figure 6.3 for a Gaussian model, the figures are similar for, e.g., Bernoulli or Poisson variables). Yet, how to guarantee this convergence theoretically is still to be understood. It might be possible to show that the associated transformation is a contraction, as it numerically seems to be the case for the ℓ^2–norm.

Conjecture 6.3. *For all exponential models $\mathcal{D}_{\exp}$, and all bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$ with a unique optimal arm, the sequence of iterates $(\tilde{\underline{v}}^{(n)})_{n\geq 0}$ defined by (6.11) converges to $\underline{w}$.*

Remark. One can state other update rules, like modifying only the weights of one sub-optimal arm together with the weight of the optimal arm, choosing this sub-optimal arm circularly or randomly. The empirical performance of those procedures is also pretty good, as far as we observed in some numerical experiments.



(a) $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$



(b) $\underline{\mu} = (0.95, 0.93, 0.92, 0.9, 0.8, 0.7, 0.5, 0.4)$

Figure 6.3: For the standard Gaussian model, coordinates of the vectors $(\underline{\tilde{v}}^{(n)})_{n\geq 0}$ with $\underline{\tilde{v}}^{(0)}$ the uniform vector; the values of the corresponding coordinates of $\underline{w}$ are dashed.

6.2.3. A New Sampling Rule for `top-two` Algorithms

We now give another point of view on the transformation $\underline{W}$, that will allow us to define a `top-two` sampling rule. We write, for all sub-optimal arms a:

$$W_a(\underline{v}) = v_a \frac{T_{\underline{v}}^{-1}}{\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})} \cdot \frac{d(\mu_a, \overline{\mu}_{a^\star,a,\underline{v}})}{d(\mu_a, \overline{\mu}_{a^\star,a,\underline{v}})} = \frac{T_{\underline{v}}^{-1}}{d(\mu_a, \overline{\mu}_{a^\star,a,\underline{v}})} \cdot \frac{v_a\, d(\mu_a, \overline{\mu}_{a^\star,a,\underline{v}})}{\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})} = \chi_a(\underline{v}) \cdot (1 - \beta_a(\underline{v}))\,,$$

where we define

$$\chi_a(\underline{v}) \overset{\text{def}}{=} \frac{T_{\underline{v}}^{-1}}{d(\mu_a, \overline{\mu}_{a^\star,a,\underline{v}})} \qquad \text{and} \qquad \beta_a(\underline{v}) \overset{\text{def}}{=} \frac{v_{a^\star}\, d(\mu^\star, \overline{\mu}_{a^\star,a,\underline{v}})}{\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})}\,. \tag{6.12}$$

$\chi_a(\underline{v})$ is the relative contribution of arm a in $T_{\underline{v}}^{-1}$, while $\beta_a(\underline{v})$ is the proportion of transportation cost $\mathrm{TC}_{a\to a^\star}(\underline{\mu},\underline{v})$ due to the transport of arm $a^\star$. We note, in passing, that vector $\underline{\chi}$ is a probability vector, which will be useful in Section 6.2.3.

Example. For the model $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ of Gaussian variables with common variance $\sigma^2 > 0$, the closed-form expression (2.21) of the Kullback-Leibler divergence gives

$$\begin{aligned}\beta_a(\underline{v}) &= \frac{v_{a^\star}\, d(\mu^\star, \overline{\mu}_{a^\star,\, a,\, \underline{v}})}{v_{a^\star}\, d(\mu^\star, \overline{\mu}_{a^\star,\, a,\, \underline{v}}) + v_a\, d(\mu_a, \overline{\mu}_{a^\star,\, a,\, \underline{v}})} \\ &= \frac{v_{a^\star}\Big(\frac{v_a}{v_a+v_{a^\star}}(\mu^\star - \mu_a)\Big)^2}{v_{a^\star}\Big(\frac{v_a}{v_a+v_{a^\star}}(\mu^\star - \mu_a)\Big)^2 + v_a\Big(\frac{v_{a^\star}}{v_a+v_{a^\star}}(\mu_a - \mu^\star)\Big)^2} \\ &= \frac{v_a}{v_a + v_{a^\star}}\,.\end{aligned}$$

We note that this quantity does not depend on $\underline{\mu}$, which is a specificity of the model.

The definition of transformation $\underline{W}$ now reads

$$\forall \underline{v} \in \mathrm{int}(\Sigma_K), \forall a \in [K], \qquad W_a(\underline{v}) = \begin{cases} \chi_a(\underline{v})(1-\beta_a(\underline{v})) & \text{if } a \neq a^\star\,, \\ \sum_{b\neq a^\star} \chi_b(\underline{v})\beta_b(\underline{v}) & \text{if } a = a^\star\,. \end{cases}$$

We note that the knowledge of $\underline{\chi}(\underline{w})$ and $\beta_a(\underline{w})$ is sufficient to generate the optimal weights $\underline{w}(\underline{\mu})$. , we select a couple of best-arm and challenger $(a^\star, a)$ where the challenger a is sampled according to the probability $\underline{\chi}(\underline{w})$. Then we choose between the optimal arm and its challenger by keeping the leader with probability $\beta_a(\underline{w})$. This results in a sample from the optimal weight vector $\underline{w}$ thanks to the fixed point property 6.2. Interestingly, we recognize here a `top-two` procedure that first samples a couple of leader and challenger and then selects an arm from this couple.

Sampling rule. To give a precise description of the `top-two` sampling rule, we need to write explicitly the dependency on $\underline{\mu}$ of the previously defined quantities $T_{\underline{v}}$, $\chi_a(\underline{v})$, $\beta_a(\underline{v})$ and $W_a(\underline{v})$, as the bandit instance is unknown to the strategy. To that end, we define $T_{\underline{v}}(\underline{\mu})$, $\chi_a(\underline{\mu}, \underline{v})$ and $W_a(\underline{\mu}, \underline{v})$ respectively by Equations (6.9), (6.12) and (6.10), and we set:

$$\forall a \neq b \in [K], \qquad \beta_{b,a}(\underline{\mu}, \underline{v}) \overset{\text{def}}{=} \begin{cases} \dfrac{v_b\, d(\mu_b, \overline{\mu}_{b,\, a,\, \underline{v}})}{v_b\, d(\mu_b, \overline{\mu}_{b,\, a,\, \underline{v}}) + v_a\, d(\mu_a, \overline{\mu}_{b,\, a,\, \underline{v}})} & \text{if } \mu_b \neq \mu_a\,, \\ \frac{1}{2} & \text{if } \mu_b = \mu_a\,, \end{cases}$$

which corresponds to the relative contribution of arm b to the transportation cost between arms a and b. Note that the dependency of $\beta_{b,a}(\underline{\mu}, \underline{v})$ on vector $\underline{v}$ is only though the ratio $x_{b,a} = \frac{v_b}{v_b+v_a}$, as

$$\beta_{b,a}(\underline{\mu}, \underline{v}) = \frac{1}{1 + \dfrac{v_a}{v_b}\dfrac{d(\mu_a, \overline{\mu}_{b,\, a,\, \underline{v}})}{d(\mu_b, \overline{\mu}_{b,\, a,\, \underline{v}})}} = \frac{1}{1 + \dfrac{1-x_{b,a}}{x_{b,a}}\dfrac{d(\mu_a, (1-x_{b,a})\mu_a + x_{b,a}\mu_b)}{d(\mu_b, (1-x_{b,a})\mu_a + x_{b,a}\mu_b)}}\,.$$

Hence we will sometimes see $\beta_{b,a}(\underline{\mu}, \cdot\,)$ as a one-variable function:

$$\forall x \in (0,1), \qquad \beta_{b,a}(\underline{\mu}, x) \overset{\text{def}}{=} \frac{1}{1 + \dfrac{1-x}{x}\dfrac{d(\mu_a, (1-x)\mu_a + x\mu_b)}{d(\mu_b, (1-x)\mu_a + x\mu_b)}}\,. \tag{6.13}$$

Once again (see, e.g., the `Track-and-Stop` strategy), at each time step t, we consider the plug-in estimates of those quantities when replacing $\underline{\mu}$ by its empirical estimate $\underline{\hat{\mu}}(t)$. We choose

Algorithm 20: $\underline{\chi}$ challenger

Input: history of observations I_{t-1}
leader L_t
Output: challenger C_t

$\underline{p} \leftarrow \underline{p}(t-1)$
$\underline{\hat{\mu}} \leftarrow \underline{\hat{\mu}}(t-1)$
for $a \neq L_t$ **do**
  $\chi_a \leftarrow \dfrac{T_{\underline{p}}^{-1}(\underline{\hat{\mu}})}{d(\hat{\mu}_a, \overline{\hat{\mu}}_{L_t,a,\underline{p}})}$
Choose C_t according to distribution $\underline{\chi}$

Algorithm 21: Adaptive `sample-arm` (sampling version)

Input: history of observations I_{t-1}
leader L_t
challenger C_t
Output: sampled arm A_t

$A_t \leftarrow \begin{cases} L_t & \text{with probability } \beta_{L_t,C_t}(\underline{\hat{\mu}}(t-1), \underline{p}(t-1)) \,, \\ C_t & \text{otherwise.} \end{cases}$

L_t as the `Empirical-Best` leader (see Algorithm 18), and as previously explained, the challenger procedure consists in[2] choosing C_t according to vector

$$\underline{\hat{\chi}}(t-1) \stackrel{\text{def}}{=} \underline{\chi}(\underline{\hat{\mu}}(t-1), \underline{p}(t-1)) \,, \qquad \text{where } \underline{p}(t-1) \stackrel{\text{def}}{=} \frac{\underline{N}(t-1)}{t-1}$$

denotes the empirical proportions of draws after time step $t-1$. The leader L_t, respectively the challenger C_t, is then pulled with probability

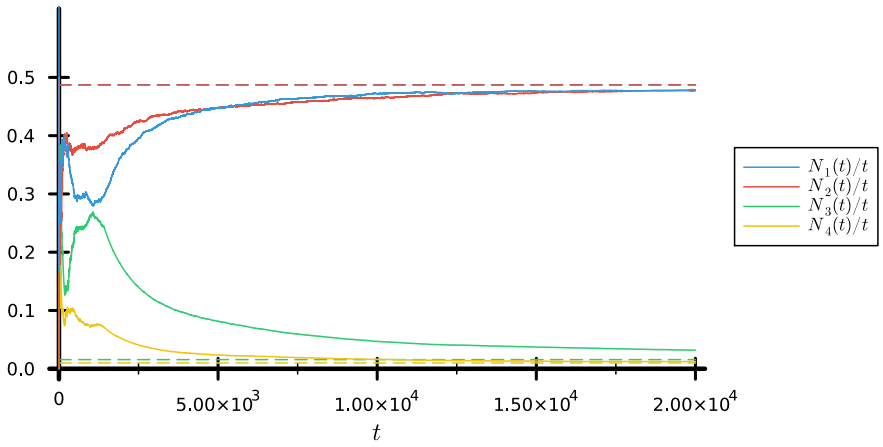
$$\hat{\beta}_{L_t,C_t}(t) \stackrel{\text{def}}{=} \beta_{L_t,C_t}(\underline{\hat{\mu}}(t-1), \underline{p}(t-1)) \,, \qquad \text{respectively } \hat{\beta}_{C_t,L_t}(t) \stackrel{\text{def}}{=} \beta_{C_t,L_t}(\underline{\hat{\mu}}(t-1), \underline{p}(t-1)) \,.$$

See Algorithms 20 and 21 for details. This adaptive strategy will be denoted by `TT-EB-`$\underline{\chi}$.

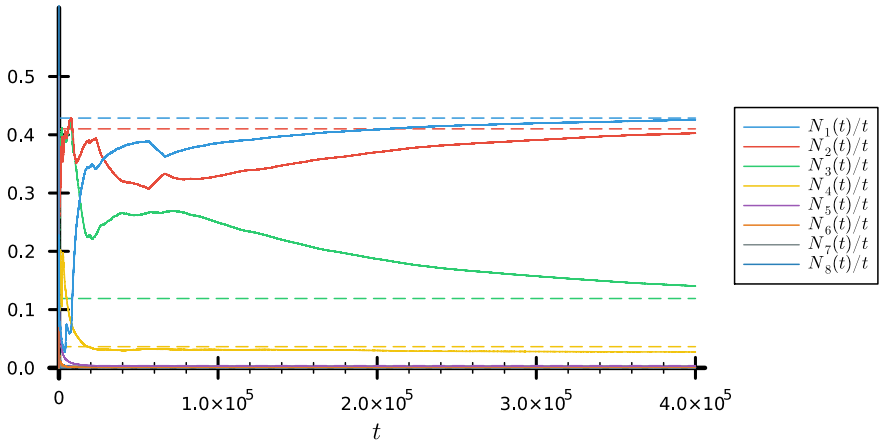
Remark. The adaptive proportions $\beta_{b,a}(\underline{\mu}, \underline{\nu})$ have also been obtained by You et al. (2023) for a general exponential model.

Experiments. The convergence of the sampling frequencies of the strategy `TT-EB-`$\underline{\chi}$ to the optimal weight vector $\underline{w}(\underline{\mu})$ is illustrated on Figures 6.4 and 6.5. Whatever the model and parameter, numerical experiments show convergence to the optimal frequencies. We observe that the evolution of the sampling frequencies is quite smooth compared to, e.g., `Track-and-Stop` or the `top-two` algorithm `TT-EB-TC` with adaptive proportions, for which the challenger is chosen as the arm minimizing the transportation cost with the leader.

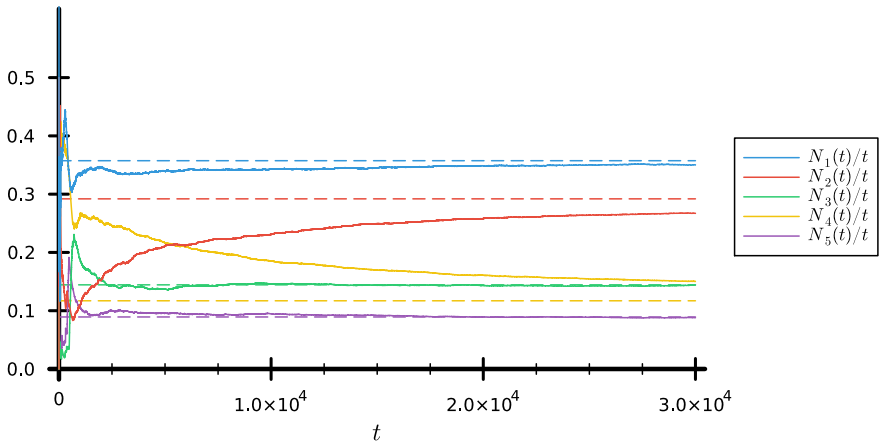
[2] To be well-defined, $\underline{\chi}$ requires $\underline{\hat{\mu}}(t-1)$ to have a unique optimal arm. If not, we can simply choose A_t among all empirically optimal arms (but note that this event occurs with null probability for a continuous model).



(a) $\underline{\mu} = (0.9, 0.7, 0.4, 0.3)$

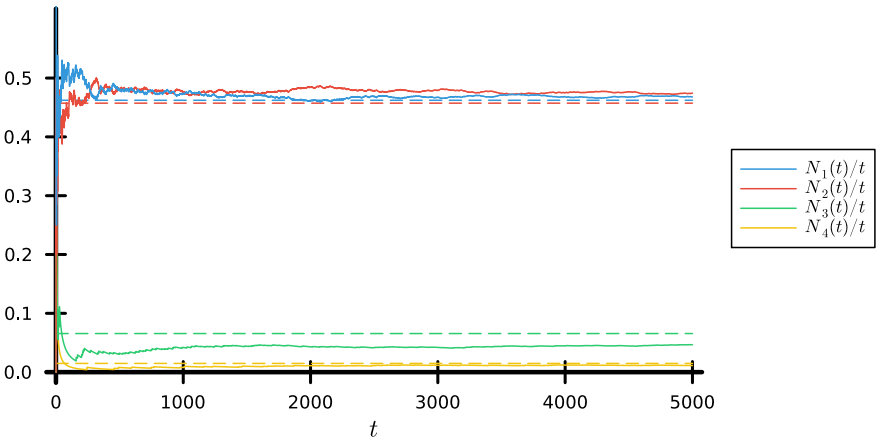


(b) $\underline{\mu} = (0.95, 0.93, 0.92, 0.9, 0.8, 0.7, 0.5, 0.4)$

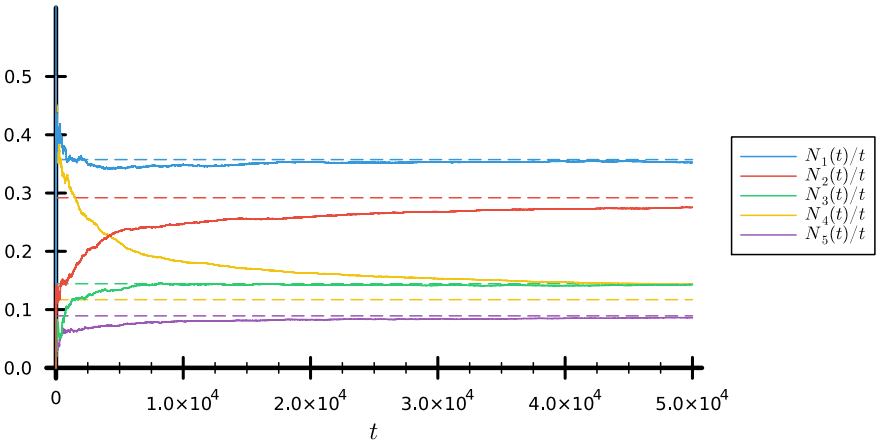


(c) $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$

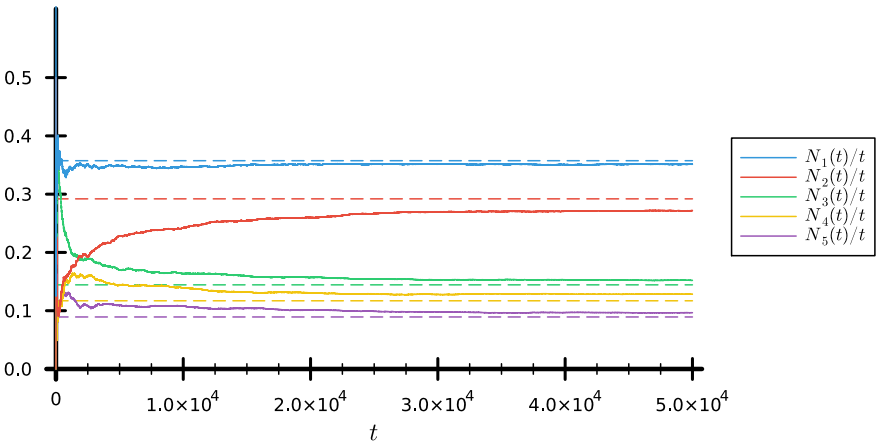
Figure 6.4: Evolution of the sampling frequencies on a simulation of TT-EB-$\underline{\chi}$ with standard Gaussian variables. The values of the corresponding coordinates of $\underline{w}(\underline{\mu})$ are dashed.



(a) Bernoulli model, $\underline{\mu} = (0.9, 0.7, 0.5, 0.1)$



(b) Bernoulli model, $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$



(c) Poisson model, $\underline{\mu} = (0.9, 0.7, 0.65, 0.63, 0.6)$

Figure 6.5: Evolution of the sampling frequencies on a simulation of TT-EB-$\underline{\chi}$ with the Bernoulli and Poisson models. The values of the corresponding coordinates of $\underline{w}(\underline{\mu})$ are dashed.

6.3. Asymptotically Optimal Adaptive Algorithms

The objective of this section is to discuss some adaptive `top-two` algorithms that could be asymptotically optimal for a general exponential model $\mathcal{D}_{\exp}$. The result was recently obtained for a Gaussian model by You et al. (2023) for the `Top-Two-Thompson-Sampling` algorithm, but their analysis also holds for, e.g., the pair of `Empirical-Best` leader and `Transportation-Cost` or `Transportation-Cost-Penalized` challenger.

Considered leaders and challengers. In order to get the simplest possible analysis, we focus in this section on the `Empirical-Best` leader given in Algorithm 18 and the `Transportation-Cost` and `Transportation-Cost-Penalized` challengers given in Algorithm 19. The adaptive versions of those `top-two` algorithms are denoted by `TT-EB-TC` and `TT-EB-TCP` (similarly to the acronyms of Jourdan et al., 2022, see also Section 2.2.7). The conjecture is given below.

Conjecture 6.4. *Let $\mathcal{D}_{\exp}$ be an exponential model. The* `TT-EB-TC` *and* `TT-EB-TCP` *algorithms with the adaptive sampling rule of Algorithm 21 satisfy, for all bandit problems $\underline{\mu}$ in $\mathcal{D}_{\exp}$ with distinct means:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu}) \,.$$

Remark. When studying `top-two` algorithms, it is a common hypothesis to work with bandits instance for which all means differ (see, e.g., Jourdan et al., 2022). We will refer to such instances as *generic* bandits. For other instances, there might be a lack of exploration, especially when the first estimates of the best arm are poor.

In the rest of this section, we provide some first arguments to prove Conjecture 6.4. We largely follow the proof structure of You et al. (2023, Appendix A) for Gaussian variables (except the additional step 2), and try to generalize their statements. As we will see, this might imply slight modifications of the considered strategies `TT-EB-TC` and `TT-EB-TCP`, which do not deeply affect the definition of the sampling rules.

Overview of the argumentation. In the sequel, we fix some generic bandit problem $\underline{\mu}$ in $\mathcal{D}_{\exp}$. Let $a^\star$ and $\underline{w}$ respectively denote its unique optimal arm and its optimal weight vector.

We split the analysis into several steps:

1. In the first step, we introduce notation and recall that asymptotic optimality can be proved by showing the convergence of the ratios $p_a(t)/p_{a^\star}(t)$ to $w_a/w_{a^\star}$ for all sub-optimal arms a,
2. Then, in step 2, we propose to slightly modify the sampling rule in order to facilitate the control of the empirical numbers of pulls.
3. To ensure the convergence of empirical estimates, we have to prove that the strategy sufficiently explores all arms. While we still do not know how to prove the sufficient exploration for `TT-EB-TC` and `TT-EB-TCP` with a general exponential model, we give a possible path of reasoning.
4. In order to prove the convergence of the ratios $p_a(t)/p_{a^\star}(t)$, You et al. (2023) showed another useful convergence in the Gaussian case. We explain in step 4 that the generalization of this convergence seems to be the difficult part of our analysis.
5. Finally, we quickly explain in step 5 how one could prove the convergence of the ratios $p_a(t)/p_{a^\star}(t)$ if the requirements of steps 3 (sufficient exploration) and 4 (supporting convergence) were satisfied.

6.3.1. Step 1: A Sufficient Condition

Let us introduce a notion of *quick convergence* for random variables.

Definition 6.5. [quick convergence]
Let $\mathbb{L}^1$ denotes the set of real valued random variables X such that $\mathbb{E}[|X|] < +\infty$.
We say that a sequence of real random variables $(X(t))_{t\in\mathbb{N}}$ *quickly converges*:

- to $x \in \mathbb{R}$, a fact which we denote $X(t) \xrightarrow[t\to+\infty]{\text{q.}} x$, if
$$\forall \varepsilon > 0, \exists T \in \mathbb{L}^1, \qquad \forall t \geq T, \quad |X(t) - x| \leq \varepsilon \,.$$
- to $+\infty$, a fact which we denote $X(t) \xrightarrow[t\to+\infty]{\text{q.}} +\infty$, if
$$\forall c > 0, \exists T \in \mathbb{L}^1, \qquad \forall t \geq T, \quad X(t) \geq c \,.$$
- to $-\infty$, a fact which we denote $X(t) \xrightarrow[t\to+\infty]{\text{q.}} -\infty$, if $-X(t) \xrightarrow[t\to+\infty]{\text{q.}} +\infty$.

We say that a sequence of random vectors $(\underline{X}(t))_{t\in\mathbb{N}}$ in $\mathbb{R}^K$ quickly converges to $\underline{x} \in \mathbb{R}^K$, a fact which we denote $\underline{X}(t) \xrightarrow[t\to+\infty]{\text{q.}} \underline{x}$, if $X_a(t) \xrightarrow[t\to+\infty]{\text{q.}} x_a$ for all $a \in [K]$.

Proving the asymptotic optimality of a strategy using the `Global-Likelihood-Ratio` stopping rule can be reduced to showing that the empirical frequencies of pulls quickly converge to $\underline{w}$, as the following result states.

Proposition 6.6. *Let $\mathcal{D}_{\text{exp}}$ be an exponential model. Consider a strategy for which the stopping rule is the* `Global-Likelihood-Ratio` *stopping rule (Algorithm 5) with threshold (6.3) for some $\alpha > 1$. If $\underline{\mu}$ is a bandit problem in $\mathcal{D}_{\mathcal{N}_{\sigma^2}}$ such that the sampling rule (applied infinitely without stopping rule) satisfies*
$$\underline{p}(t) \xrightarrow[t\to+\infty]{\text{q.}} \underline{w} \,,$$
then
$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu}) \,.$$

This result was originally stated for a Gaussian model by Qin et al. (2017), but the generalization to exponential models is straightforward[3].

In fact, the quick convergence of $\underline{p}(t)$ to $\underline{w}$ is equivalent to the quick convergence of the ratios $(p_a(t)/p_{a^\star}(t))_{a\neq a^\star}$:
$$\underline{p}(t) \xrightarrow[t\to+\infty]{\text{q.}} \underline{w} \qquad \Longleftrightarrow \qquad \forall a \neq a^\star, \quad \frac{p_a(t)}{p_{a^\star}(t)} \xrightarrow[t\to+\infty]{\text{q.}} \frac{w_a}{w_{a^\star}} \,.$$
This equivalence is easy to prove, using that $\underline{p}(t)$ and $\underline{w}$ belong to Σ_K for the indirect sense. Hence
$$\forall a \neq a^\star, \quad \frac{p_a(t)}{p_{a^\star}(t)} \xrightarrow[t\to+\infty]{\text{q.}} \frac{w_a}{w_{a^\star}} \qquad \Longrightarrow \qquad \limsup_{\delta\to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log 1/\delta} \leq T(\underline{\mu}) \,.$$

Objective. To prove the conjecture, we now focus on demonstrating that
$$\forall a \neq a^\star, \qquad \frac{p_a(t)}{p_{a^\star}(t)} \xrightarrow[t\to+\infty]{\text{q.}} \frac{w_a}{w_{a^\star}} \,. \tag{6.14}$$

[3] And even to more general models, up to technicalities, see Jourdan et al. (2022, Theorem 2 in Appendix C.5).

Remark. You et al. (2023) actually worked with a stronger condition than quick convergence: instead of considering random variables in $\mathbb{L}^1$, they require those variables to be in $\cap_{p\geq 1}\mathbb{L}^p$, following tools developed by Qin and Russo (2022). Yet, quick convergence is sufficient for what we want to prove.

6.3.2. Step 2: Tracking versus Sampling

Due to the randomization of the strategy (and especially the random choice between the leader and the challenger), the vector $\underline{p}(t)$ might largely deviate from its expected value. When working with (sub-)Gaussian variables, this deviation might be controlled so that it does not really complicate the arguments (see You et al., 2023). With a general exponential model, we might ease the analysis by replacing the sampling with a tracking: no concentration result will be required. The technique was considered by Jourdan and Degenne (2023) to obtain non-asymptotic guarantees for a Top-Two algorithm.

Tracking. When choosing between the leader and the challenger, one can replace the sampling choice with a C-tracking of the targeted proportions. We now explain the modification in detail.

We will need to define tracking procedures associated with fixed pairs of leader and challenger. To that end, for a pair of arms $a \neq b$ and a time step t, we consider the event

$$\mathcal{C}^{a,b}(t) \stackrel{\text{def}}{=} \Big\{ L_t = a \text{ and } C_t = b \Big\},$$

which indicates a time step at which the leader and challenger are respectively arms a and b. We count the cumulative number of pulls of arms a and b under those events and denote them by:

$$N_a^{a,b}(t) \stackrel{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = a\}\, \mathbb{I}\Big\{\mathcal{C}^{a,b}(s)\Big\} \qquad \text{and} \qquad N_b^{a,b}(t) \stackrel{\text{def}}{=} \sum_{s\in[t]} \mathbb{I}\{A_s = b\}\, \mathbb{I}\Big\{\mathcal{C}^{a,b}(s)\Big\}\,,$$

and we also define the average target proportions of arms a and b under those events:

$$\overline{\beta}_a^{a,b}(t) \stackrel{\text{def}}{=} \frac{1}{N^{a,b}(t)} \sum_{s\in[t]} \hat{\beta}_{a,b}(s) \mathbb{I}\Big\{\mathcal{C}^{a,b}(s)\Big\}\,,$$

$$\text{and} \qquad \overline{\beta}_b^{a,b}(t) \stackrel{\text{def}}{=} \frac{1}{N^{a,b}(t)} \sum_{s\in[t]} \hat{\beta}_{b,a}(s) \mathbb{I}\Big\{\mathcal{C}^{a,b}(s)\Big\} = 1 - \overline{\beta}_a^{a,b}(t)\,,$$

$$\text{where} \qquad N^{a,b}(t) \stackrel{\text{def}}{=} N_a^{a,b}(t) + N_b^{a,b}(t) = \sum_{s\in[t]} \mathbb{I}\Big\{\mathcal{C}^{a,b}(s)\Big\}\,.$$

We modify the choice of arm A_t between the leader L_t and the challenger C_t so as to track the associated average target proportions, see Algorithm 22.

Algorithm 22: Adaptive `sample-arm` (tracking version)

Input: history of observations I_{t-1}
leader L_t
challenger C_t
Output: sampled arm A_t

$$A_t \leftarrow \begin{cases} L_t & \text{if } \dfrac{N_{L_t}^{L_t,C_t}(t-1)}{N_{L_t}^{L_t,C_t}(t-1) + N_{C_t}^{C_t,L_t}(t-1)} < \overline{\beta}_{L_t}^{L_t,C_t}(t)\,, \\ C_t & \text{otherwise.} \end{cases}$$

With that modification, we get precise control of the quantities $N_a^{a,b}(t)$ and $N_b^{a,b}(t)$:

Lemma 6.7. *For all pairs (a,b) of arms with $a \neq b$, and $t \geq 0$, we get*

$$\left|N_a^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_a^{a,b}(t)\right| = \left|N_b^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_b^{a,b}(t)\right| \leq 1\,.$$

Before proving this result, we explain why it might be useful. Given a good choice of leader, we hope that $L_t = a^\star$ except for a sub-linear number of time steps. If this property holds, then

$$\forall a \neq a^\star, \qquad p_a(t) \simeq \frac{N_a^{a^\star,a}(t)}{t} \simeq \frac{N^{a^\star,a}(t)}{t}\,\overline{\beta}_a^{a^\star,a}(t)\,,$$

and proving the condition (6.14) (or maybe $\underline{p}(t) \xrightarrow[t\to+\infty]{\text{q.}} \underline{w}$) might be done by studying the asymptotic behaviour of $N^{a^\star,a}(t)/t$ and $\overline{\beta}_a^{a^\star,a}(t)$.

Proof. The equality comes from the relationship:

$$N_b^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_{b,a}(t) = N^{a,b}(t) - N_a^{a,b}(t) - N^{a,b}(t)(1-\overline{\beta}_{a,b}(t)) = N^{a,b}(t)\,\overline{\beta}_{a,b}(t) - N_a^{a,b}(t)\,.$$

To obtain the inequality, we proceed by induction on t. Let us assume that the result holds at time step $t-1$. If at time step t, we have $L_t \neq a$ or $C_t \neq b$, then all considered quantities are not updated and the result holds by induction. Now, if $L_t = a$ and $C_t = b$, let us assume that arm $A_t = a$ (one can adapt the argument when $A_t = b$) and prove that

$$-1 \leq N_a^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_a^{a,b}(t) \leq 1\,. \tag{6.15}$$

By definition of A_t (see Algorithm 22), we know, as $A_t = a$, that

$$N_a^{a,b}(t-1) \leq N^{a,b}(t-1)\,\overline{\beta}_a^{a,b}(t)\,,$$

hence, using that $N_a^{a,b}(t) = N_a^{a,b}(t-1) + 1$ and $N^{a,b}(t) = N^{a,b}(t-1) + 1$:

$$N_a^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_a^{a,b}(t) = N_a^{a,b}(t-1) + 1 - N^{a,b}(t-1)\,\overline{\beta}_a^{a,b}(t) - \overline{\beta}_a^{a,b}(t) \leq 1\,,$$

which gives the second inequality of (6.15). For the first inequality, remark that

$$N^{a,b}(t)\,\overline{\beta}_a^{a,b}(t) = N^{a,b}(t-1)\,\overline{\beta}_a^{a,b}(t-1) + \beta_{a,b}(t)\,,$$

which, together with the induction property, leads to

$$N_a^{a,b}(t) - N^{a,b}(t)\,\overline{\beta}_a^{a,b}(t) = \underbrace{N_a^{a,b}(t-1) + 1 - N^{a,b}(t-1)\,\overline{\beta}_a^{a,b}(t-1)}_{\geq 0 \text{ by induction}} - \beta_{a,b}(t) \geq -1\,.$$

This concludes the proof. □

6.3.3. Step 3: Sufficient Exploration

A typical methodology in fixed-confidence best-arm identification (see, e.g., the use of forced exploration for `Track-and-Stop` in page 49) is to ensure that empirical quantities converge to their expected values by requiring that the number of pulls of arms diverges almost surely, with a minimal sub-linear rate that we take (arbitrarily) of order $\sqrt{t}$.

Definition 6.8. [sufficient exploration]
We say that a strategy provides sufficient exploration with rate $\sqrt{t}$ if

$$\exists C > 0, \exists T \in \mathbb{L}^1, \qquad \forall t \geq T, \quad \min_{a \in [K]} N_a(t) \geq C\sqrt{t}\,. \tag{6.16}$$

Jourdan et al. (2022) proved that both TT-EB-TC-β and TT-EB-TCP-β, which are non-adaptive versions of TT-EB-TC and TT-EB-TCP, provide sufficient exploration for an exponential model. Their result does not generalize to adaptive versions for which the probability of choosing the leader instead of the challenger might be arbitrarily close to 0 or 1. You et al. (2023) explained how to override this limitation in the Gaussian case. In the last paragraph of this section, we will explain why we do think that the result still holds for an exponential model, without providing explicit proof.

Forced exploration. Meanwhile, we can use a tool that ensures that sufficient exploration is provided: we might simply force exploration (like, e.g., the Track-and-Stop strategy) and hence assume that this step is satisfied for free. We recall the process quickly (see also page 49). We define, at a given time step t, the set $U(t-1)$ of arms that are under-sampled with respect to the sub-linear rate $\sqrt{t}$:

$$U(t-1) \overset{\text{def}}{=} \left\{ a \in [K] \,:\, N_a(t-1) < \sqrt{\frac{t}{K}} + 1 \right\}.$$

If this set is not empty, we overrule the choice of A_t given by the top-two sampling rule and replace it by picking the least pulled arm. This implies that the following lemma holds.

Lemma 6.9. *The* TT-EB-TC *and* TT-EB-TCP *algorithms, with the forced exploration process defined above, provide sufficient exploration with rate* $\sqrt{t}$*:*

$$\forall t \geq 0, \qquad \min_{a \in [K]} N_a(t) \geq \sqrt{\frac{t}{K}}\,.$$

Remark.
- With forced exploration, we do not have to restrict our attention to a generic bandit problem $\underline{\mu}$. The analysis stands for all bandit problems with a unique optimal arm.
- With forced exploration, the penalization of the Transportation-Cost-Penalized challenger is useless. Yet, as we aim to prove that forced exploration is not necessary at least for TT-EB-TCP, we still consider this strategy in this section.

Convergence of the empirical means. The sufficient exploration property implies an important property which is the quick convergence of the empirical means to $\underline{\mu}$.

Proposition 6.10. *Consider a strategy for which sufficient exploration holds with rate* $\sqrt{t}$*. Then*

$$\underline{\hat{\mu}}(t) \quad \underset{t \to +\infty}{\overset{q.}{\longrightarrow}} \quad \underline{\mu}\,.$$

Remark. This result can be used, together with Proposition 6.6 to prove the asymptotic optimality of the Track-and-Stop and Exploration-Biased-Sampling algorithms, as we detail in Section 4.

The proof is an adaptation of Garivier and Kaufmann (2016, Lemma 19).

Proof. Let $C > 0$ and $T \in \mathbb{L}^1$ satisfying (6.16). Fix $\varepsilon > 0$ and define, for $t \geq 1$:

$$\mathcal{E}_t = \bigcap_{s \geq t} \bigcap_{a \in [K]} \Big\{ |\hat{\mu}_a(t) - \mu_a| \leq \varepsilon \Big\} .$$

We want to prove that

$$\mathbb{E}_{\underline{\mu}}[T_\varepsilon] < +\infty , \qquad \text{where } T_\varepsilon \overset{\text{def}}{=} \inf\Big\{ t \geq 1 : \mathcal{E}_t \Big\} . \tag{6.17}$$

We get

$$\begin{aligned}
\mathbb{E}_{\underline{\mu}}[T_\varepsilon] &= \sum_{t \geq 0} \mathbb{P}_{\underline{\mu}}(T_\varepsilon > t) \\
&= \sum_{t \geq 0} \mathbb{P}_{\underline{\mu}}\Big(T_\varepsilon > t, T > t\Big) + \mathbb{P}_{\underline{\mu}}\Big(T_\varepsilon > t, t \geq T\Big) \\
&= \sum_{t \geq 0} \mathbb{P}_{\underline{\mu}}\Big(T > t\Big) + \mathbb{P}_{\underline{\mu}}\Big(T_\varepsilon > t, t \geq T\Big) \\
&= \mathbb{E}_{\underline{\mu}}[T] + \sum_{t \geq 0} \mathbb{P}_{\underline{\mu}}\Big(\mathcal{E}_t^{\mathsf{c}}, t \geq T\Big)
\end{aligned} \tag{6.18}$$

Fix $t \geq 0$. We have, using optional skipping and the definition of t

$$\begin{aligned}
\mathbb{P}_{\underline{\mu}}\Big(\mathcal{E}_t^{\mathsf{c}}, t \geq T\Big) &= \mathbb{P}_{\underline{\mu}}\Big(\exists s \geq t, \exists a \in [K], |\hat{\mu}_a(s) - \mu_a| > \varepsilon, t \geq T\Big) \\
&\leq \mathbb{P}_{\underline{\mu}}\Big(\exists n \geq C\sqrt{t}, \exists a \in [K], |\hat{\mu}_{a,n} - \mu_a| > \varepsilon, t \geq T\Big) \\
&\leq \sum_{a \in [K]} \sum_{n \geq C\sqrt{t}} \mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_{a,n} - \mu_a| > \varepsilon, t \geq T\Big) \\
&\leq \sum_{a \in [K]} \sum_{n \geq C\sqrt{t}} \mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_{a,n} - \mu_a| > \varepsilon\Big) .
\end{aligned}$$

For $a \in [K]$ and $n \geq 0$, a Cramér-Chernoff bound gives

$$\begin{aligned}
\mathbb{P}_{\underline{\mu}}\Big(|\hat{\mu}_{a,n} - \mu_a| > \varepsilon\Big) &= \mathbb{P}_{\underline{\mu}}\Big(\hat{\mu}_{a,n} < \mu_a - \varepsilon\Big) + \mathbb{P}_{\underline{\mu}}\Big(\hat{\mu}_{a,n} > \mu_a + \varepsilon\Big) \\
&\leq \exp\Big(-n d(\mu_a - \varepsilon, \mu_a)\Big) + \exp\Big(-n d(\mu_a + \varepsilon, \mu_a)\Big) .
\end{aligned}$$

Hence we get that

$$\begin{aligned}
\mathbb{P}_{\underline{\mu}}\Big(\mathcal{E}_t^{\mathsf{c}}, t \geq T\Big) &\leq \sum_{a \in [K]} \frac{\exp\Big(-C\sqrt{t} d(\mu_a - \varepsilon, \mu_a)\Big)}{1 - \exp\Big(-d(\mu_a - \varepsilon, \mu_a)\Big)} + \frac{\exp\Big(-C\sqrt{t} d(\mu_a + \varepsilon, \mu_a)\Big)}{1 - \exp\Big(-d(\mu_a + \varepsilon, \mu_a)\Big)} \\
&\leq 2KC_1 \exp(-C_2\sqrt{t}) ,
\end{aligned} \tag{6.19}$$

where we set

$$C_1 \overset{\text{def}}{=} \max_{a \in [K]} \left(\max\left(\frac{1}{1 - \exp(-d(\mu_a - \varepsilon, \mu_a))}, \frac{1}{1 - \exp(-d(\mu_a + \varepsilon, \mu_a))} \right) \right) < +\infty ,$$

$$\text{and} \quad C_2 \overset{\text{def}}{=} C \cdot \min_{a \in [K]} \Big(\min\big(d(\mu_a - \varepsilon, \mu_a), d(\mu_a + \varepsilon, \mu_a) \big) \Big) > 0 .$$

Injecting (6.19) into (6.18) finally leads to $\mathbb{E}_{\underline{\mu}}[T_\varepsilon] < +\infty$ and concludes the proof of (6.17). □

Avoiding forced exploration. We hope to avoid the use of forced exploration with additional arguments. Especially since this sufficient exploration occurs for non-adaptive versions of `TT-EB-TC` and `TT-EB-TCP`, and it seems that the adaptive versions are kind of encouraging exploration: given a pair (ℓ, c) of leader and challenger, pulling arm ℓ (respectively c) will increase the probability of pulling c (respectively ℓ) at next time step the pair of leader and challenger will be the same. This is a consequence of the monotonicity of the function $x \in (0,1) \mapsto \beta_{\ell,c}(\underline{\mu}, x)$ defined in (6.13), where we recall that x is interpreted as the proportion $\frac{v_\ell}{v_\ell + v_c}$ associated to a weight vector $\underline{v}$.

Lemma 6.11. *Let $a, b \in [K]$ be such that $\mu_a \neq \mu_b$. Then $\beta_{b,a}(\underline{\mu}, \cdot)$ is a decreasing (continuous) function such that*

$$\lim_{x \to 1} \beta_{b,a}(\underline{\mu}, x) = 0\,, \qquad \textit{and} \qquad \lim_{x \to 0} \beta_{b,a}(\underline{\mu}, x) = 1\,.$$

Proof. The continuity of $\beta_{b,a}(\underline{\mu}, \cdot)$ is a direct consequence of the continuity properties of d (see page 43), since

$$\forall x \in (0,1), \qquad \beta_{b,a}(\underline{\mu}, x) \overset{\text{def}}{=} \frac{1}{1 + \dfrac{1-x}{x} \dfrac{d(\mu_a, (1-x)\mu_a + x\mu_b)}{d(\mu_b, (1-x)\mu_a + x\mu_b)}}$$

To prove the decreasing of $\beta_{b,a}(\underline{\mu}, \cdot)$, we use the strict convexity of d on its domain. Let $0 < x < x' < 1$. By setting $\overline{\mu}_x = (1-x)\mu_a + x\mu_b$, we get $\overline{\mu}_x \in (\mu_a, \overline{\mu}_{x'})$ or $\overline{\mu}_x \in (\overline{\mu}_{x'}, \mu_a)$, depending on the relative positions of μ_a and μ_b, hence there exists $\alpha \in (0,1)$ such that

$$\overline{\mu}_x = \alpha \overline{\mu}_{x'} + (1-\alpha)\mu_a\,.$$

In fact, it is easy to see that $\alpha = \frac{x}{x'}$. By strict convexity of d, we get that

$$d(\mu_a, \overline{\mu}_x) \leq \alpha d(\mu_a, \overline{\mu}_{x'}) + (1-\alpha) \underbrace{d(\mu_a, \mu_a)}_{=0} = \frac{x}{x'} d(\mu_a, \overline{\mu}_{x'})\,,$$

hence we proved that

$$\frac{d(\mu_a, \overline{\mu}_x)}{x} < \frac{d(\mu_a, \overline{\mu}_{x'})}{x'}\,.$$

Similarly, one can show that

$$\frac{d(\mu_b, \overline{\mu}_x)}{1-x} > \frac{d(\mu_b, \overline{\mu}_{x'})}{1-x'}\,,$$

which entails that $\beta_{b,a}(\underline{\mu}, x) > \beta_{b,a}(\underline{\mu}, x')$. This concludes the proof of the decreasing of $\beta_{b,a}(\underline{\mu}, \cdot)$.

We now prove the limit behaviors. We recall the expression of the mean-parameterized Kullback-Leibler divergence of the model defined in Equation (6.4):

$$\forall \mu, \mu' \in \mathcal{M}, \qquad d(\mu, \mu') \overset{\text{def}}{=} \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = (\theta - \theta') b'(\theta) - b(\theta) + b(\theta')\,,$$

where $\theta = (b')^{-1}(\mu)$ and $\theta' = (b')^{-1}(\mu')$. As b is twice differentiable, a first-order Taylor expansion of $d(\mu\,, \cdot\,)$ ensures that for all $\mu \in \mathcal{M}$,

$$d(\mu, \mu') = o_{\mu' \to \mu}(\mu' - \mu)\,.$$

Injecting this asymptotic behaviour in the expression (6.13) of $\beta_{b,a}(\underline{\mu}, x)$ leads to

$$\begin{aligned}\beta_{b,a}(\underline{\mu}, x) &= \frac{1}{1 + \dfrac{1-x}{x}\dfrac{d(\mu_a, (1-x)\mu_a + x\mu_b)}{d(\mu_b, (1-x)\mu_a + x\mu_b)}} \\ &= \frac{1}{1 + \dfrac{1-x}{x}\dfrac{o_{x\to 0}(x(\mu_b - \mu_a))}{d(\mu_b, (1-x)\mu_a + x\mu_b)}} \\ &= \frac{1}{1 + \dfrac{(\mu_b - \mu_a)}{d(\mu_b, \mu_a)} o_{x\to 0}(1)} \quad \xrightarrow[x\to 0]{} 1\,.\end{aligned}$$

The second result can be derived similarly. □

6.3.4. Step 4: A Useful Relationship

To obtain the required convergence (6.14), an important argument in the final step will be to use that, if there exists $a \neq a^\star$ such that $p_a(t)/p_{a^\star}(t)$ is a little bit above its target $w_a/w_{a^\star}$, then if t is large enough, there exists another sub-optimal arm b such that $p_b(t)/p_{a^\star}(t)$ is slightly under $w_b/w_{a^\star}$. This property occurs, for instance (see the next step for precise calculations), as soon as

$$\sum_{a\neq a^\star} \frac{p_a(t)}{p_{a^\star}(t)} \quad \xrightarrow[t\to+\infty]{\text{q.}} \quad \sum_{a\neq a^\star} \frac{w_a}{w_{a^\star}}\,. \tag{6.20}$$

In the Gaussian case, You et al. (2023) were able to prove that

$$\sum_{a\neq a^\star} \frac{p_a(t)^2}{p_{a^\star}(t)^2} \quad \xrightarrow[t\to+\infty]{\text{q.}} \quad 1 = \sum_{a\neq a^\star} \frac{w_a^2}{w_{a^\star}^2}\,,$$

and uses this relation to get the desired property. Yet, this clearly uses the fact that the optimal weight vector $\underline{w}$ satisfies a property that is highly specific to Gaussian variables:

$$w_{a^\star}^2 = \sum_{a\neq a^\star} w_a^2\,,$$

that can be obtained directly from Equation (6.8) or by the analysis of Chapter 3, see Equation (3.12).

It seems that this Gaussian argument is the most challenging to extend the analysis of You et al. (2023) to all exponential models. A general relation that might be obtained to replace this Gaussian property is that the frequencies of pulls satisfy the following quick convergence:

$$\sum_{b\neq a^\star} \frac{d(\mu^\star, \overline{\mu}_{a^\star,b,\underline{p}(t)})}{d(\mu_b, \overline{\mu}_{a^\star,b,\underline{p}(t)})} \quad \xrightarrow[t\to+\infty]{\text{q.}} \quad 1 = \sum_{b\neq a^\star} \frac{d(\mu^\star, \overline{\mu}_{a^\star,b,\underline{w}})}{d(\mu_b, \overline{\mu}_{a^\star,b,\underline{w}})}\,,$$

where the equality stands from Equation (6.8).

$p_{a^\star}(t)$ bounded away from 0. An important implication of such convergence is that the ratio $p_a(t)/p_{a^\star}(t)$ does not blow up, i.e., that $p_{a^\star}(t)$ is bounded away from 0 after a sufficiently large enough time:

$$\exists p_{\min} > 0, \exists T \in \mathbb{L}^1, \qquad \forall t \geq T, \quad p_{a^\star}(t) \geq p_{\min}\,.$$

If another proof structure is required to replace this step, it might be possible to obtain this property[4] by using Lemma 6.11, which, as previously discussed, indicates that the strategy somewhat encourages exploration.

[4] Even without forced exploration.

6.3.5. Step 5: Quick Convergence of $p_a(t)/p_{a^\star}(t)$

In this last step we assume that the convergence (6.20) is satisfied:

$$\sum_{a \neq a^\star} \frac{p_a(t)}{p_{a^\star}(t)} \xrightarrow[t\to+\infty]{\text{q.}} \sum_{a \neq a^\star} \frac{w_a}{w_{a^\star}} \,. \tag{6.21}$$

In this step, we prove (6.14) with similar arguments to You et al. (2023). Some simplifications are made by the use of the deterministic challenger rules considered in this work (`Transportation-Cost` and `Transportation-Cost-Penalized`) instead of the Bayesian challenger of You et al. (2023).

Lemma 6.12. *For all $\varepsilon > 0$, there exists $T \in \mathbb{L}^1$ such that for all $t \geq T$ and $a \neq a^\star$,*

$$\frac{p_a(t)}{p_{a^\star}(t)} \geq \frac{w_a + \varepsilon}{w_{a^\star}} \qquad \Longrightarrow \qquad A_{t+1} \neq a \,.$$

This result states that, after a sufficiently long time, if arm a is over-pulled at time step t, it has no chance to be pulled at $t+1$.

For the sake of completeness, we will give a proof of that result that was already known (see, e.g., Jourdan et al., 2022, Page 7). We need to introduce notation and preliminary results. To simplify the reading, we set:

$$\forall t \geq 0, \forall a \neq b, \qquad \widehat{\mathrm{TC}}_{a\to b}(t) \stackrel{\text{def}}{=} \mathrm{TC}_{a\to b}(\hat{\underline{\mu}}(t), \underline{p}(t))$$

For $a \neq a^\star$, we also define

$$f_a(x) \stackrel{\text{def}}{=} d\Big(\mu^\star, \frac{x}{1+x}\mu_a + \frac{1}{1+x}\mu^\star\Big) + x\, d\Big(\mu_a, \frac{x}{1+x}\mu_a + \frac{1}{1+x}\mu^\star\Big) \,,$$

and note that

$$\mathrm{TC}_{a\to a^\star}(\underline{\mu}, \underline{v}) = v_{a^\star} f_a\Big(\frac{v_a}{v_{a^\star}}\Big) \,,$$

As a consequence of Proposition 6.1, we have that

$$\forall a \neq a^\star, \qquad f_a\Big(\frac{w_a}{w_{a^\star}}\Big) = \frac{T(\underline{\mu})^{-1}}{w_{a^\star}} \,. \tag{6.22}$$

We will use the fact that f_a is an increasing function.

Lemma 6.13. **[Garivier and Kaufmann, 2016, Appendix 5.2]**
For all $a \neq a^\star$, $x \in (0, +\infty) \mapsto f_a(x)$ is an increasing function.

We also use the sufficient exploration property (Lemma 6.9) to ensure that the empirical costs are quite close to their theoretical value after a sufficiently long time:

Lemma 6.14. *For all $\eta > 0$, there exists $T \in \mathbb{L}^1$ such that*

$$\forall t \geq T, \forall a \neq a^\star, \quad (1-\eta)\widehat{\mathrm{TC}}_{a\to a^\star}(t) \leq \mathrm{TC}_{a\to a^\star}(\underline{\mu}, \underline{p}(t)) \leq (1+\eta)\widehat{\mathrm{TC}}_{a\to a^\star}(t) \,.$$

Proof. This is a consequence of Proposition 6.10 and the continuity of the transportation costs. □

Proof of Lemma 6.12. For the sake of simplicity, we only prove the result for the `Transportation-Cost` challenger. The same proof applies for the `Transportation-Cost-Penalized` challenger, up to slightly modifying Lemma 6.14 with the penalized empirical costs.

Let $\varepsilon > 0$. Thanks to the previous steps, we get the existence of some integrable times satisfying the following properties:

- By assumption (6.21), there exists $T_1 \in \mathbb{L}^1$ such that

$$\forall t \geq T_1, \qquad \left| \sum_{a \neq a^\star} \frac{p_a(t)}{p_{a^\star}(t)} - \sum_{a \neq a^\star} \frac{w_a}{w_{a^\star}} \right| \leq \frac{\varepsilon}{w_{a^\star}} \,. \tag{6.23}$$

- In addition, note that by Proposition 6.10, we get the existence of $T_2 \in \mathbb{L}^1$ such that for all $t \geq T_2$

$$\forall t \geq T_2, \forall a \in [K], \qquad |\hat{\mu}_a(t) - \mu_a| < \frac{\Delta_{\min}}{2} \,,$$

  where $\Delta_{\min} = \min_{a \neq a^\star} \Delta_a$. As a consequence $L_t = a^\star$ for all $t \geq T_2$.
- Also, by Lemma 6.14, for a fixed $\eta > 0$ there exists $T_3 \in \mathbb{L}^1$ such that for all $t \geq T_3$,

$$\forall a \neq a^\star, \qquad (1-\eta)\,\mathrm{TC}_{a \to a^\star}(t) \leq \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{p}(t)) \leq (1+\eta)\,\mathrm{TC}_{a \to a^\star}(t) \,. \tag{6.24}$$

We now set $T \overset{\text{def}}{=} \max(T_1, T_2, T_3) \in \mathbb{L}^1$ where T_3 is associated with a value of η to be chosen later.

Let $t \geq T$ and assume that there exists $a \neq a^\star$ such that

$$\frac{p_a(t)}{p_{a^\star}(t)} \geq \frac{w_a + \varepsilon}{w_{a^\star}} \,. \tag{6.25}$$

We will prove that $A_{t+1} \neq a$. As $L_{t+1} = a^\star$, we only need to prove that $C_{t+1} \neq a$. By definition of C_{t+1} (which is an arm minimizing its empirical transportation cost with $a^\star$), it suffices to prove that

$$\widehat{\mathrm{TC}}_{b \to a^\star}(t) < \widehat{\mathrm{TC}}_{a \to a^\star}(t) \,. \tag{6.26}$$

By Equation (6.23), the condition (6.25) implies that

$$\exists b \neq a^\star, \qquad \frac{p_b(t)}{p_{a^\star}(t)} \leq \frac{w_b}{w_{a^\star}} \,. \tag{6.27}$$

Using the monotonicity property of f_a and f_b (Lemma 6.13), this leads to

$$\begin{aligned}
(1+\eta)\widehat{\mathrm{TC}}_{a \to a^\star}(t) &\geq \mathrm{TC}_{a \to a^\star}(\underline{\mu}, \underline{p}(t)) && \text{by (6.24)} \\
&= p_{a^\star}(t) \cdot f_a\Big(\frac{p_a(t)}{p_{a^\star}(t)}\Big) \\
&\geq p_{a^\star}(t) \cdot f_a\Big(\frac{w_a + \varepsilon}{w_{a^\star}}\Big) && \text{by (6.25)} \\
&= p_{a^\star}(t) \cdot (1+\varepsilon_a) \cdot f_a\Big(\frac{w_a}{w_{a^\star}}\Big) \\
&= p_{a^\star}(t) \cdot (1+\varepsilon_a) \cdot f_b\Big(\frac{w_b}{w_{a^\star}}\Big) && \text{by (6.22)} \\
&\geq (1+\varepsilon_a) \cdot p_{a^\star}(t) \cdot f_b\Big(\frac{p_b(t)}{p_{a^\star}(t)}\Big) && \text{by (6.27)} \\
&= (1+\varepsilon_a) \cdot \mathrm{TC}_{b \to a^\star}(\underline{\mu}, \underline{p}(t)) && \text{by (6.24)} \\
&\geq (1+\varepsilon_a)(1-\eta)\widehat{\mathrm{TC}}_{b \to a^\star}(t) \,,
\end{aligned}$$

where we defined

$$\varepsilon_a \overset{\text{def}}{=} \frac{f_a\left(\frac{w_a+\varepsilon}{w_{a^\star}}\right)}{f_a\left(\frac{w_a}{w_{a^\star}}\right)} - 1\,,$$

which is positive by Lemma 6.13. Hence we proved that

$$\widehat{\mathrm{TC}}_{a\to a^\star}(t) \geq \frac{(1-\eta)}{(1+\eta)}(1+\varepsilon_a)\widehat{\mathrm{TC}}_{b\to a^\star}(t)\,.$$

Taking η small enough, such that

$$\frac{(1-\eta)}{(1+\eta)}\Big(1+\min_{a\neq a^\star}\varepsilon_a\Big) > 1\,,$$

leads to inequality (6.26), which enforces $C_t \neq a$ and completes the proof. □

Using Lemma 6.12, it is not hard to prove the convergence of $p_a(t)/p_{a^\star}(t)$. The analysis of You et al. (2023) might be followed, by proving first that, for all $\varepsilon > 0$, there exists $T \in \mathbb{L}^1$ such that

$$\forall t \geq T, \forall a \neq a^\star, \qquad \frac{p_a(t)}{p_{a^\star}(t)} \leq \frac{w_a+\varepsilon}{w_{a^\star}}\,,$$

and then deduce the required objective (6.14) by using assumption (6.21).

6.3.6. Conclusion

In this section, we proposed some ideas to generalize the Gaussian analysis of You et al. (2023) and prove Conjecture 6.4. In steps 2 and 3, we slightly modified the definition of the strategies `TT-EB-TC` and `TT-EB-TCP` in order to simplify their analysis for an exponential model. In step 5, we proved that the Gaussian arguments still apply for those models, provided that some hypothetical condition holds. It seems that the last biggest challenge is to prove that this condition is satisfied (see discussions in step 4).

About the $\underline{\chi}$ challenger. The analysis made in this section might be extended to the $\underline{\chi}$ challenger presented in Section 6.2.3, but the randomization of the challenger complicates the analysis as well as technical issues with non-continuous models. For this challenger, a first step might be to obtain the convergence of the empirical frequencies of pulls with the knowledge of $\underline{\mu}$, stated in the following conjecture.

Conjecture 6.15. *Consider a version of the* `TT-EB-`$\underline{\chi}$ *algorithm which knows the values of* $\underline{\mu}$ *and use it instead of* $\underline{\hat{\mu}}(t-1)$ *to compute* $\underline{\hat{\chi}}(t-1)$ *and* $\widehat{\beta}_{L_t,C_t}(t)$*. Then*

$$\underline{p}(t) \underset{t\to+\infty}{\longrightarrow} \underline{w} \qquad \mathbb{P}_{\underline{\mu}}\text{–a.s.}\,.$$

6.4. Side Note: On the Asymptotic Optimality of `Track-and-Stop`

In this section, we give a proof of the asymptotic optimality of `Track-and-Stop` . When introducing their strategy, Garivier and Kaufmann (2016) proved that `Track-and-Stop` satisfies the following asymptotic bound, which does not exactly correspond to asymptotic optimality in the sense of Definition 2.12.

Theorem 6.16. **[Garivier and Kaufmann, 2016, Theorem 14]**
Consider an exponential model $\mathcal{D}_{\exp}$. The `Track-and-Stop` *strategy, with threshold* (6.3) *for a fixed $\alpha > 1$, satisfies*

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\exp}, \qquad \limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq \alpha T(\underline{\mu}) \,.$$

As a by-product of some of the results stated in Section 6.4, we show that `Track-and-Stop` is asymptotically optimal whatever the choice of $\alpha > 1$.

Theorem 6.17. *Consider an exponential model $\mathcal{D}_{\exp}$. The* `Track-and-Stop` *strategy, with C-tracking and the* `Global-Likelihood-Ratio` *stopping rule with threshold* (6.3) *for a fixed $\alpha > 1$, is asymptotically optimal:*

$$\forall \underline{\mu} \text{ in } \mathcal{D}_{\exp}, \qquad \limsup_{\delta \to 0} \frac{\mathbb{E}_{\underline{\mu}}[\tau_\delta]}{\log \frac{1}{\delta}} \leq T(\underline{\mu}) \,.$$

Remark. The proposition also applies for the `Exploration-Biased-Sampling` strategy studied in Chapter 4. One can follow the same proof structure, recalling that this strategy naturally has sufficient exploration with rate $\sqrt{t}$.

The result is deduced from the general Propositions 6.6 and 6.10.

Proof. We will prove that the hypothesis of Proposition 6.6 is satisfied, i.e., that

$$\underline{p}(t) \quad \xrightarrow[t \to +\infty]{\text{q.}} \quad \underline{w} \,.$$

We rely on the following lemma, which controls the values of the number of pulls.

Lemma 6.18. **[Garivier and Kaufmann, 2016, Lemma 7]**
The `Track-and-Stop` *algorithm with C-tracking satisfies, for all $t \geq 0$ and $a \in [K]$,*

$$N_a(t) \geq \sqrt{t + K^2} - 2K \,, \tag{6.28}$$

$$\text{and} \quad \left| p_a(t) - \frac{1}{t} \sum_{s \in [t]} \underline{\hat{w}}(s) \right| \leq K \frac{1 + \sqrt{t}}{t} \,. \tag{6.29}$$

We deduce from (6.28) that the strategy provides sufficient exploration with rate $\sqrt{t}$, for some constant $C \in (0, 1)$. By Proposition 6.10, this implies that

$$\underline{\hat{\mu}}(t) \quad \xrightarrow[t \to +\infty]{\text{q.}} \quad \underline{\mu} \,,$$

and by continuity of $\underline{\hat{\mu}} \mapsto \underline{w}(\underline{\hat{\mu}})$ at $\underline{\mu}$, that

$$\underline{\hat{w}}(t) = \underline{w}(\underline{\hat{\mu}}(t)) \quad \xrightarrow[t \to +\infty]{\text{q.}} \quad \underline{w} \,.$$

It can be easily seen that this implies the quick convergence of the Cesàro sum:

$$\frac{1}{t} \sum_{s \in [t]} \underline{\hat{w}}(s) \quad \xrightarrow[t \to +\infty]{\text{q.}} \quad \underline{w} \,.$$

By Equation (6.29), this implies that

$$\underline{p}(t) \quad \xrightarrow[t\to+\infty]{\mathsf{q}\cdot} \quad \underline{w}\,,$$

and concludes the proof. □

6.5. Conclusion

In this chapter, we first considered a new transformation of weights which is of high interest for the sample complexity optimization problem, as its unique fixed point is the optimal weight vector. We derived procedures for the computation of the optimal weight vector that revealed empirically correct and efficient, but we are still looking for theoretical arguments that to guarantee the convergence.

Then, we came to grips with adaptive `top-two` algorithms in order to prove their asymptotic optimality for a general exponential model. Following the Gaussian proof structure of You et al. (2023), we showed that some arguments might easily be generalized or bypassed (in a preliminary study, by using, e.g., forced exploration). However, a crucial convergence argument presented in step 4 is missing in the general case, and constitutes the main remaining challenge, unless different proof structures avoiding this argument are studied. Indeed, in future work, we might also investigate another path based on the three following steps:

- prove Conjecture 6.3, that is, prove the convergence of the empirical procedures proposed in Section 6.2.2 for the computation of $\underline{w}(\underline{\mu})$ *with* the knowledge of $\underline{\mu}$,
- independently, prove that the `TT-EB-TC` and/or `TT-EB-TCP` strategies have sufficient exploration, using the partial ideas introduced in Section 6.3.3,
- finally, prove the asymptotic optimality by combining the results of the two first phases.

Bibliography

S. Agrawal, S. Juneja, and P. Glynn. Optimal δ–Correct Best-Arm Selection for Heavy-Tailed Distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 61–110. PMLR, 2020.

S.M. Ali and S.D Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society*, 28:131–142, 1966.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-Armed Bandits. In *Proceedings of the 23th Conference on Learning Theory*, 2010.

A. Barrier, A. Garivier, and T. Kocák. A Non-Asymptotic Approach to Best-Arm Identification for Gaussian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022.

A. Barrier, A. Garivier, and G. Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 136–181. PMLR, 2023.

R.G. Bartle and D.R. Sherbert. *Introduction to Real Analysis*. John Wiley & Sons, 3rd edition, 2000.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

J. Bretagnolle and C. Huber. Estimation des densités : risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47, 1979.

A.N. Burnetas and M.N. Katehakis. Optimal Adaptive Policies for Sequential Allocation Problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

A. Carpentier and A. Locatelli. Tight (Lower) Bounds for the Fixed Budget Best Arm Identification Bandit Problem. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pages 590–604. PMLR, 2016.

L. Chen, J. Li, and M. Qiao. Towards Instance Optimal Bounds for Best Arm Identification. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 535–592. PMLR, 2017.

H. Chernoff. Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

Y. Chow and H. Teicher. *Probability Theory*. Springer, 1988.

S.-R. Chowdhury, P. Saux, O.-A. Maillard, and A. Gopalan. Bregman Deviations of Generic Exponential Families. In *Proceedings of the 36th Conference on Learning Theory*, volume 195, pages 394–449. PMLR, 2023.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

R. Degenne. On the Existence of a Complexity in Fixed Budget Bandit Identification. In *Proceedings of the 36th Conference on Learning Theory*, volume 195, pages 1131–1154. PMLR, 2023.

R. Degenne and W.M. Koolen. Pure Exploration with Multiple Correct Answers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

R. Degenne, W.M. Koolen, and P. Ménard. Non-Asymptotic Pure Exploration by Solving Games. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J.L. Doob. *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons, 1953.

E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7 (39):1079–1105, 2006.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

A. Garivier and É. Kaufmann. Optimal Best Arm Identification with Fixed Confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pages 998–1027. PMLR, 2016.

A. Garivier and É. Kaufmann. Non-Asymptotic Sequential Tests for Overlapping Hypotheses Applied to Near-Optimal Arm Identification in Bandit Models. *Sequential Analysis*, 40(1):61–96, 2021.

A. Garivier, P. Ménard, and G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

A. Garivier, H. Hadiji, P. Ménard, and G. Stoltz. KL-UCB-switch: Optimal Regret Bounds for Stochastic Bandits from Both a Distribution-Dependent and a Distribution-Free Viewpoints. *Journal of Machine Learning Research*, 23(179):1–66, 2022.

G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3rd edition, 2001.

J. Honda and A. Takemura. Non-Asymptotic Analysis of a New Bandit Algorithm for Semi-Bounded Rewards. *Journal of Machine Learning Research*, 16(113):3721–3756, 2015.

L.J. Hong, W. Fan, and J. Luo. Review on Ranking and Selection: A New Perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

S.R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 423–439. PMLR, 2014.

M. Jourdan and R. Degenne. Non-Asymptotic Analysis of a UCB-Based Top Two Algorithm, 2023. Preprint, arXiv:2210.05431.

M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and É. Kaufmann. Top Two Algorithms Revisited. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC Subset Selection in Stochastic Multi-armed Bandits. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Z. Karnin, T. Koren, and O. Somekh. Almost Optimal Exploration in Multi-Armed Bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1238–1246. PMLR, 2013.

M. Kato, K. Ariu, M. Imaizumi, M. Nomura, and C. Qin. Optimal Best Arm Identification in Two-Armed Bandits with a Fixed Budget under a Small Gap, 2022. Preprint, arXiv:2201.04469.

É. Kaufmann and S. Kalyanakrishnan. Information Complexity in Bandit Subset Selection. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 228–251. PMLR, 2013.

É. Kaufmann and W.M. Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

É. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

T. Kocák and A. Garivier. Best Arm Identification in Spectral Bandits. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, volume 3, pages 2220–2226. International Joint Conferences on Artificial Intelligence Organization, 2020.

J. Komiyama. Suboptimal Performance of the Bayes Optimal Algorithm in Frequentist Best Arm Identification, 2022. Preprint, arXiv:2202.05193.

J. Komiyama, T. Tsuchiya, and J. Honda. Minimax Optimal Algorithms for Fixed-Budget Best Arm Identification. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

T.L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 2nd edition, 1998.

S. Mannor and J.N. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

P. Ménard. Gradient Ascent for Active Exploration in Bandit Problems, 2019. Preprint, arXiv:1905.08165.

C. Qin and D. Russo. Electronic Companion to Adaptivity and Confounding in Multi-Armed Bandit Experiments, 2022. SSRN:4115833.

C. Qin, D. Klabjan, and D. Russo. Improving the Expected Improvement Algorithm. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

D. Russo. Simple Bayesian Algorithms for Best Arm Identification. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pages 1417–1418. PMLR, 2016.

D. Russo. Simple Bayesian Algorithms for Best Arm Identification. *Operations Research*, 68(6): 1625–1647, 2020.

X. Shang, R. de Heide, É. Kaufmann, P. Ménard, and M. Valko. Fixed-Confidence Guarantees for Bayesian Best-Arm Identification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1823–1832. PMLR, 2020.

M. Simchowitz, K. Jamieson, and B. Recht. The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1794–1834. PMLR, 2017.

G. Stoltz. Lecture Notes on "Learning & Sequential Optimization", 2022.

W.R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast Pure Exploration via Frank-Wolfe. In *Advances in Neural Information Processing Systems*, volume 34, pages 5810–5821. Curran Associates, Inc., 2021.

W. You, C. Qin, Z. Wang, and S. Yang. Information-Directed Selection for Top-Two Algorithms. In *Proceedings of the 36th Conference on Learning Theory*, volume 195, pages 2850–2851. PMLR, 2023.

Index

Contributions à une théorie de l'exploration pure en statistique séquentielle

Résumé. Cette thèse, à la croisée entre les domaines de l'intelligence artificielle, de la statistique séquentielle et de l'optimisation, s'intéresse au problème d'identification du meilleur bras (en espérance) dans les bandits non structurés à K bras. Ce problème possède deux approches dont les niveaux de compréhension sont très différents.

Le cadre à confiance fixée est le mieux compris : des stratégies asymptotiquement optimales sont connues, et l'on s'intéresse à l'obtention de garanties non asymptotiques pour des stratégies (si possible) simples et naturelles. Avec des bandits Gaussiens, nous proposons l'analyse à risque fini d'une nouvelle stratégie (asymptotiquement optimale) grâce aux propriétés de régularité de ce modèle. Cette stratégie modifie subtilement la règle d'attribution des tirages de l'algorithme Track-and-Stop en une règle plus prudente et interprétable. Dans le contexte plus général d'un modèle exponentiel, nous proposons l'ébauche d'une analyse de l'asymptotique optimalité d'algorithmes de type Top-Two adaptatifs, dont les règles de choix de tirages sont particulièrement simples.

Par ailleurs, dans le cadre à budget fixé, où l'existence d'une hypothétique complexité reste à démontrer, nous proposons des généralisations à des modèles non-paramétriques des bornes (supérieures et inférieures) connues jusqu'à présent pour des modèles très spécifiques. Les bornes obtenues font intervenir des quantités de théorie de l'information plus précises que les écarts entre les moyennes qui apparaissaient précédemment. Ces nouvelles quantités pourraient être la clé pour mesurer la complexité de l'identification de meilleur bras à budget fixé.

Mots-clés. Problèmes de bandits · Identification de meilleur bras · Statistiques séquentielles · Apprentissage statistique · Intelligence artificielle

---

Contributions to a Theory of Pure Exploration in Sequential Statistics

Abstract. This thesis lies in the fields of artificial intelligence, sequential statistics and optimization. We focus on the problem of best (in expectation) arm identification in unstructured muti-armed bandits. This problem has two approaches with very different levels of understanding.

The fixed-confidence framework is the best understood: asymptotically optimal strategies are known, and we are interested in obtaining non-asymptotic guarantees for (if possible) simple and natural strategies. Working with Gaussian bandits, we propose a finite risk analysis of a new (asymptotically optimal) strategy using the regularity properties of this model. This strategy slightly modifies the sampling rule of the Track-and-Stop algorithm into a more conservative and interpretable rule. In the more general context of an exponential model, we propose a preliminary analysis of the asymptotic optimality of adaptive Top-Two algorithms, whose sampling rules are particularly simple.

Independently, in the fixed-budget framework, for which the existence of a hypothetical complexity remains to be demonstrated, we propose generalizations to non-parametric models of the existing bounds (upper and lower) that were available so far only for very specific models. The obtained bounds involve more precise information-theoretic quantities than the gaps (differences between the means) which appeared previously. These new quantities could be the key to measuring the complexity of fixed-budget best-arm identification.

Keywords. Multi-Armed Bandits · Best-Arm Identification · Sequential Statistics · Statistical Learning · Machine Learning