On the notion of optimality in the stochastic multi-armed bandit problems

Pierre Ménard

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *3 Juillet 2018* **par :**

Pierre Ménard

Sur la notion d'optimalité dans les problèmes de bandit stochastique

---

JURY

Philippe Berthet	Université Toulouse 3	Examinateur
Alexandra Carpentier	Université de Magdeburg	Rapporteur
Aurélien Garivier	Université Toulouse 3	Directeur de thèse
Tor Lattimore	DeepMind Londres	Examinateur
Béatrice Laurent-Bonneau	INSA de Toulouse	Examinateur
Gilles Stoltz	CNRS-Université Paris Sud	Directeur de thèse
Michal Valko	Inria Lille - Nord Europe	Rapporteur

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :
Institut de Mathématiques de Toulouse (UMR 5219)

Directeur(s) de Thèse :
Aurélien Garivier et *Gilles Stoltz*

Rapporteurs :
Alexandra Carpentier et *Michal Valko*

Remerciements

Plutôt que d'arguer combien l'écriture de ces remerciements est un exercice ô combien délicat et périlleux, je tiens à t'avertir, toi lecteur, qui en ce moment même se lance frénétiquement à la recherche de ton prénom pendant que je me ridiculise au tableau. Ne perds pas espoir trop vite, car dans le cas hautement improbable où après une lecture attentive de ces quelques lignes, tu ne trouverais pas les remerciements (sûrement mérités) auxquels tu t'attendais, je te propose, au lieu de te lancer dans un autodafé avec tous les exemplaires de ce manuscrit se trouvant à ta portée, de compléter les remerciements interactif ci-dessous et ainsi réparer l'immense affront que j'aurais pu commettre.

Je tiens en premier lieu à remercier mes deux directeurs de thèse, Aurélien Garivier et Gilles Stoltz. Ce fut un plaisir de travailler avec vous pendant ces trois années ; vous m'avez laissé faire ce qui m'intéressait tout en évitant que je ne me perde. J'ai appris beaucoup de choses à vos cotés. En particulier vous m'avez fait découvrir un domaine qui me passionne ; Aurélien, je pense que je ne pourrais pas respecter ton interdiction de travailler sur les bandits stochastiques à K bras dans les deux années qui viennent. Je vous remercie aussi pour votre expertise, rigueur et disponibilité, Aurélien, malgré ton emploi du temps plus que chargé, Gilles ; nous nous sommes un peu moins vu ces deux dernières années, mais cela ne nous a pas empêché de travailler ensemble, d'autant plus que tu es une des rares personnes (l'unique ?) à répondre à mes mails dans l'heure, peu importe le moment de l'envoie. Merci également pour m'avoir plusieurs fois accueilli à Paris[1].

Je voudrais remercier mes deux autres co-auteurs : un petit frère de thèse, Hédi Hadiji, et un grand frère de thèse, Sébastien Gerchinovitz, j'espère que nous collaborerons de nouveau dans un avenir proche. Sébastien, discuter de mathématiques avec toi est toujours un plaisir et j'espère que le nouveau souffle que tu as insufflé au groupe de lecture, avec en particulier l'inclusion des doctorants (j'espère y avoir contribué un peu aussi) perdurera.

Je souhaite aussi remercier Alexandra Carpentier et Michal Valko pour avoir rapporté cette thèse ainsi que Philippe Berthet, Tor Lattimore et Béatrice Laurent-Bonneau pour avoir accepté de faire partie de mon jury. Alexandra, merci également de m'avoir accueilli un mois à Magdeburg (sans oublier Andrea, Géraud, Claire et Elizabeth !) .

[1]Gilles, je sais que tu trouveras ces remerciements beaucoup trop brefs mais sache que je garde un souvenir impérissable de ton premier retour sur mon rapport de stage noirci de commentaires.

Je me dois aussi de remercier les membres et ex-membres du Bureau 207 qui ont dû me supporter pendant plus ou moins longtemps : Claire D. qui a tout de suite décelé mon talent pour les percussions et, dès lors, n'a eu de cesse de m'encourager dans cette voie, Kevin pour ces discussions sur le cinéma (coréen en particulier), Ioana pour ton salto arrière départ chaise de bureau (et aussi pour notre passion commune des arrivées matinales), Phuong pour sa bonne humeur indéfectible, Paula pour sa discrétion, Eva pour ta passion dévorante du plateau de Saclay (je te confie l'héritage du bureau 207 sans craintes puisque j'ai l'impression que tu y es depuis plus longtemps que moi).

Je ne vais pas oublier de remercier les autres membres de l'IMT que j'ai eu la chance de croiser durant cette thèse (ou avant) : Valentin mon moussin qui m'a fait découvrir le padel, Maelysse (ça s'écrit comme ça non ?), égérie à ses dépens du groupe Giphy, Mickael[2], Fabien M. : bonne continuation, Antoine "Gloire à notre président bien aimé", William (ça te dit on va à A-B...), Hugo pour ces marches aléatoires en Kangoo[3], Guillaume grand inquisiteur de l'huile de palme et des gobelets en plastiques, Laure (merci d'avoir relu ces remerciements !), Kamila pour m'avoir accompagné lors de mes NOMBREUSES formations, Kevin l'algébriste pour toutes ces studieuses préparations de TD à la mécanique des fluides, Silvère (en revanche arrête de me répéter que tu veux retourner enseigner à l'IUT), Clément #Piyavskii, Anton auteur inégalé de mails pour le séminaire doctorant ("Latina est, non legitur"), Valentin le stagiaire, Léonard (bon on n'a pas gagné...), Baptiste, Jbar (tu es parti trop tôt), François, Tristan, Magali, Camille, Jose, Maël, Thien, Sofiane pour ses subtils dessins, Stéphane, Jonathan.

Il me faut aussi remercier les membres du "groupe", messieurs Le Faou (et aussi Svetlana) kaggle master de la médiane, Huré (et son fidèle destrier hakuna matata sans doute dévoré par les mouches à l'heure où j'écris ces mots) et Quentin ("j'ai faim").

Il me reste à remercier les membres de ma famille qui ont toujours été là pour moi : Papa, Maman, Mimi, Anne (j'attends avec impatience de lire tes remerciements), Fabien (bon, il faudra que j'attende un peu plus pour les tiens).

Je tiens à remercier, du fond du coeur, ..., pour
.....
.....
Encore une fois MERCI !

[2]fondateur de l'UT4 Mickael Albertus, directeur du pôle raclette/fondu de l'IMT et tennisman à ses heures perdues.

[3]qui a démontré à la surprise générale que le 142 rue Bonnat était un état récurrent.

Contents

Abstract

The topics addressed in this thesis lie in statistical machine learning and sequential statistic. Our main framework is the stochastic multi-armed bandit problems. In this work we revisit lower bounds on the regret. We obtain non-asymptotic, distribution-dependent bounds and provide simple proofs based only on well-known properties of Kullback-Leibler divergence. These bounds show in particular that in the initial phase the regret grows almost linearly, and that the well-known logarithmic growth of the regret only holds in a final phase. Then, we propose algorithms for regret minimization in stochastic bandit models with exponential families of distributions or with distribution only assumed to be supported by the unit interval, that are simultaneously asymptotically optimal (in the sense of Lai and Robbins lower bound) and minimax optimal. We also analyze the sample complexity of sequentially identifying the distribution whose expectation is the closest to some given threshold, with and without the assumption that the mean values of the distributions are increasing. This work is motivated by phase I clinical trials, a practically important setting where the arm means are increasing by nature. Finally we extend Fano's inequality, which controls the average probability of (disjoint) events in terms of the average of some Kullback-Leibler divergences, to work with arbitrary unit-valued random variables. Several novel applications are provided, in which the consideration of random variables is particularly handy. The most important applications deal with the problem of Bayesian posterior concentration (minimax or distribution-dependent) rates and with a lower bound on the regret in non-stochastic sequential learning.

keywords: Stochastic multi-armed bandits, information theory, non-asymptotic lower bounds, regret analysis, upper confidence bound (UCB), minimax optimality, asymptotic optimality, thresholding bandits, best arm identification, unimodal regression, multiple-hypotheses testing.

Résumé

Cette thèse s'inscrit dans les domaines de l'apprentissage statistique et de la statistique séquentielle. Le cadre principal est celui des problèmes de bandit stochastique à plusieurs bras. Dans une première partie, on commence par revisiter les bornes inférieures sur le regret. On obtient ainsi des bornes non-asymptotiques dépendantes de la distribution que l'on prouve de manière très simple en se limitant à quelques propriétés bien connues de la divergence de Kullback-Leibler. Puis, on propose des algorithmes pour la minimisation du regret dans les problèmes de bandit stochastique paramétrique dont les bras appartiennent à une certaine famille exponentielle ou non-paramétrique en supposant seulement que les bras sont à support dans l'intervalle unité, pour lesquels on prouve l'optimalité asymptotique (au sens de la borne inférieure de Lai et Robbins) et l'optimalité minimax. On analyse aussi la complexité pour l'échantillonnage séquentielle visant à identifier la distribution ayant la moyenne la plus proche d'un seuil fixé, avec ou sans l'hypothèse que les moyennes des bras forment une suite croissante. Ce travail est motivé par l'étude des essais cliniques de phase I, où l'hypothèse de croissance est naturelle. Finalement, on étend l'inégalité de Fano qui contrôle la probabilité d'évènements disjoints avec une moyenne de divergences de Kullback-leibler à des variables aléatoires arbitraires bornées sur l'intervalle unité. Plusieurs nouvelles applications en découlent, les plus importantes étant une borne inférieure sur la vitesse de concentration de l'a posteriori Bayésien et une borne inférieure sur le regret pour un problème de bandit non-stochastique.

mots-clés : Bandits stochastiques multi-bras, théorie de l'information, bornes inférieures non-asymptotiques, analyse du regret, optimalité asymptotique, optimalité minimax, borne supérieure de confiance, bandits à seuil, identification du meilleur bras, régression unimodale, test d'hypothèses multiples.

Avant-propos

Cette thèse s'inscrit dans les domaines de l'*apprentissage statistique* et de la *statistique séquentielle*. Plus précisément nous nous intéresserons aux *problèmes de bandit à plusieurs bras* qui peuvent se décrire comme des problèmes d'*allocation séquentielle de ressources* dans un environnement inconnu : un agent est confronté à une collection d'alternatives inconnues, il doit alors répartir séquentiellement les essais d'alternative qui lui sont alloués afin de maximiser un certain objectif. Le paradigme récurrent est d'imaginer un agent devant une collection de bandits manchots, justifiant ainsi l'appellation. Chacune de ces machines distribue une récompense selon un certain processus inconnu de l'agent, certaines machines étant plus rentables que d'autres. À chaque tour, il tire un des bras, i.e., joue sur une des machines, et reçoit la récompense associée. Un objectif peut être alors, par exemple, de maximiser ses gains cumulés.

Il est possible de reformuler avec ce cadre théorique nombre d'autres problèmes issus de l'optimisation, l'apprentissage par renforcement ou l'apprentissage en ligne. Quant aux applications pratiques, elles s'étendent des essais cliniques (la motivation initiale des problèmes de bandit, voir Thompson [1933]) aux heuristiques d'exploration pour la résolution de jeux (voir par exemple Silver et al. [2016]). Deux grandes classes de problèmes de bandit se distinguent selon la modélisation du processus délivrant les récompenses adoptée. D'un coté les problèmes de bandit *adversarial*, voir Auer et al. [2002b], où un adversaire décide de la récompense à attribuer à chacun des bras. De l'autre, les problèmes de bandit *stochastique* où les récompenses sont issues d'un certain modèle statistique. Cette dernière peut encore être scindée en deux avec d'une part une approche bayésienne de la modélisation, de l'évaluation et de la résolution des problèmes de bandit (voir Gittins [1979] et Gittins et al. [2011]) et d'autre part, une approche plutôt fréquentiste. Il faut cependant nuancer cette séparation un peu stricte. Il n'est pas rare qu'un algorithme issu de l'une de ces deux approches soit analysé en suivant la seconde, comme par exemple l'algorithme de Thompson Sampling (voir Thompson [1933] et Korda et al. [2013]).

Les principaux outils utilisés pour traiter ces problèmes sont, pour caricaturer, des *inégalités de déviations* pour les bornes supérieures et des *inégalités d'information* pour les bornes inférieures sur la quantité d'intérêt, même si ces deux outils ne sont pas sans liens. Le plus souvent, il n'est pas suffisant de se contenter des inégalités génériques présentent dans la littérature, notamment à cause de l'aspect *séquentielle* des problèmes étudiés. Il faut alors développer des inégalités ad hoc telles que des inégalités de dévia-

tions auto-normalisées.

Dans un premier temps, Chapitre 1, on introduira les problèmes de bandit et l'on présentera les différents résultats de cette thèse. Dans les Chapitres 2, 3 et 4, on s'attachera à étudier le *regret*. C'est le critère historique associé aux problèmes de bandit qui fait apparaître le célèbre *dilemme exploration-exploitation*. On peut voir ce dernier comme l'écart entre les gains qu'un agent aurait pu obtenir en connaissant à l'avance le processus qui délivre les récompenses et ce que l'agent a réellement obtenu. Dans le Chapitre 2 on exhibera plusieurs bornes inférieures sur le regret permettant de décrire différents régimes de croissance du regret ainsi que des preuves simplifiées de bornes inférieures existantes. Puis dans le Chapitre 3 on s'intéressera à un cadre paramétrique : on supposera que les récompenses sont issues d'une certaine famille exponentielle. On y présentera un algorithme simultanément *asymptotiquement optimal* (la notion d'optimalité pour la minimisation du regret historiquement étudiée) et *minimax optimal* (une seconde notion d'optimalité inspirée des problèmes de bandit adversarial). Puis dans le Chapitre 4 on généralisera les résultats obtenus au chapitre précédent à un cadre non paramétrique où l'on supposera seulement que les récompenses sont bornées.

Dans le Chapitre 5 on cherchera à identifier le bras plus proche d'un seuil donné, cela le plus efficacement possible, avec ou sans l'hypothèse que les moyennes des bras forment une suite croissante. C'est une alternative à la minimisation du regret motivée par l'étude des essais cliniques de phase I.

Enfin, dans le Chapitre 6, un peu à l'écart des chapitres précédents, on présentera des variations autour de l'inégalité de Fano et les bornes inférieures que l'on peut en déduire. On présentera notamment deux applications : une borne inferieure sur la vitesse de concentration du posterior Bayésien et une sur le regret pour un problème de bandit adversarial avec des pertes creuses. Les outils utilisés sont similaires à ceux employés dans le Chapitre 2, c'est pourquoi nous avons choisi de l'inclure dans cette thèse.

Chapter 1

Introduction générale

Contents

1.1 Bandit stochastique paramétrique

Commençons par présenter le problème de *bandit stochastique* introduit par Thompson dans l'article fondateur Thompson [1933] puis étudié par Lai et ses co-auteurs, voir entre autres Lai and Robbins [1985] et Lai [1987]. Un problème de bandit $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ est une collection de K *bras* chacun de ces bras étant une certaine distribution. Pour la première partie de cette introduction, on se restreindra, par souci de clarté et de simplicité, à des bras Bernoulli $\nu_a = \mathrm{Ber}(\mu_a)$ de moyenne $\mu_a \in [0,1]$.

La procédure se déroule de la façon suivante : à chaque tour $1 \leqslant t \leqslant T$ l'agent tire un bras $A_t \in \{1,\dots,K\}$ puis reçoit et observe une récompense Y_t distribuée selon le bras ν_{A_t} conditionnellement indépendante du passé.

Une stratégie ψ, adoptée par l'agent, associe un bras à l'information récoltée durant les tours précédents, et éventuellement un aléa auxiliaire, qui, sans perte de généralité, peut être donné par une suite $U_0, U_1, U_2, \dots$ de variables aléatoires indépendantes, distribuées selon la loi uniforme sur $[0,1]$. Ces variables sont aussi indépendantes des récompenses Y_t. Ainsi, une stratégie est une suite de fonctions mesurables $\psi = (\psi_t)_{t\geqslant 0}$ qui à l'information passée

$$I_t = \big(U_0,\, Y_1, U_1,\, \dots,\, Y_t, U_t\big)\,,$$

associent un bras $\psi_t(I_t) = A_{t+1} \in \{1,\dots,K\}$, où $t \geqslant 0$. L'information initiale se réduit alors à $I_1 = U_0$ et le premier bras est tiré selon $A_1 = \psi_0(U_0)$. On dira que la stratégie est déterministe lorsqu'elle ne dépend pas de l'aléa auxiliaire $U_0, U_1, U_2, \dots$.

Un objectif pour l'agent peut être de maximiser l'espérance de ses gains

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right].$$

Cela revient, de manière équivalente, en notant

$$\mu^\star = \max_{a=1,\dots,K} \mu_a$$

la moyenne des bras optimaux, à minimiser le *regret cumulé* (que l'on confondra avec le regret)

$$R_{T,\underline{\nu},\psi} = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right].$$

Lorsque ce sera clair d'après le contexte on ne précisera pas la dépendance en le problème de bandit $\underline{\nu}$ ou en la stratégie ψ. Le regret correspond à l'écart entre le gain cumulé moyen qu'un agent aurait pu obtenir s'il connaissait à l'avance les moyennes des bras et celui que l'agent a réellement obtenu. Parfois il est plus utile de réécrire le regret en faisant intervenir les *écarts* entre les moyennes des bras et la plus grande des moyennes

$$\Delta_a = \mu^\star - \mu_a \quad \text{pour } a \in \{1,\dots,K\}.$$

On dira alors qu'un bras a est *sous-optimal* si $\Delta_a > 0$. En effet, en conditionnant, il est facile de voir que

$$\begin{aligned} R_T &= \sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{E}\big[\mathbb{E}[\mu^\star - Y_t | I_{t-1}]\mathbb{1}_{\{A_t=a\}}\big] \\ &= \sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{E}[\Delta_a \mathbb{1}_{\{A_t=a\}}] \\ &= \sum_{a=1}^{K} \Delta_a \mathbb{E}\big[N_a(T)\big]\,, \end{aligned}$$

où $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s=a\}}$ est le nombre de fois que l'agent a tiré le bras a jusqu'à l'instant t. On a utilisé pour la seconde égalité que conditionnellement à I_{t-1}, la récompense Y_t est une réalisation indépendante du passé de la loi ν_{A_t}. Cette formulation permet de montrer que le regret croît, même pour un mauvais algorithme, au pire linéairement avec l'horizon T.

1.1.1 Algorithme UCB

Une première approche naïve pour minimiser le regret est de tirer à chaque tour le bras ayant la plus grande moyenne empirique courante. Soit la stratégie : tirer chaque bras une fois puis, pour $t \geqslant K$:

$$A_{t+1} \in \arg\max_{a=1,\dots,K} \widehat{\mu}_a(t),$$

où les moyennes empiriques sont données par

$$\widehat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \mathbb{1}_{\{A_s=a\}} Y_s.$$

Ici l'agent se contente d'*exploiter* l'information qu'il a récoltée. Néanmoins cette stratégie peut s'avérer inefficace. En effet si, par exemple, $K = 2$, $\mu_1 = 1/2$ et $\mu_2 = \varepsilon$ où $1/2 > \varepsilon$, avec probabilité $\varepsilon/2$ on observe 0 pour le bras 1 au premier tour puis 1 pour le bras 2 au second. Alors l'agent tirera uniquement le bras 2 par la suite, et donc le regret aura une croissance linéaire en T

$$R_T \geqslant \frac{\varepsilon}{2}\left(\frac{1}{2} - \varepsilon\right)(T-1).$$

Ce qui comme nous l'avons vu correspond au pire des cas. On aurait pu se douter que quelque chose n'allait pas en remarquant que l'on compare des moyennes empiriques issues d'échantillons de tailles différentes, $N_a(t)$ ici. Un moyen pour rendre ces comparaisons plus équitable est de sous-échantillonner un des bras. Par exemple si $K = 2$, avec $N_1(t) \geqslant N_2(t)$, on sous-échantillonne le bras 1 pour obtenir un échantillon taille $N_2(t)$ dont on calcule la moyenne empirique $\widetilde{\mu}_1(t)$. Puis l'on tire le bras ayant la plus grande

moyenne empirique entre $\widetilde{\mu}_1(t)$ et $\widehat{\mu}_2(t)$. On vient de décrire en substance l'algorithme BESA (Best Empirical Sampled Average) de Baransi et al. [2014].

À l'opposé, on peut essayer d'*explorer* le plus possible en tirant uniformément au hasard, à chaque tour t, un bras A_t parmi $\{1, \dots, K\}$. Mais le regret est de nouveau linéaire avec cette nouvelle stratégie

$$R_T = \sum_{a=1}^{T} \Delta_a \frac{T}{K}.$$

Il faut donc trouver un compromis entre *exploration* et *exploitation*. Une première solution très simple consiste à utiliser l'algorithme ε-greedy, voir Sutton and Barto [1998], alternant entre exploration et exploitation avec un certain ratio. On fixe $0 < \varepsilon < 1$ puis à chaque tour t, on joue avec probabilité $1-\varepsilon$ un bras $A_{t+1} \in \arg\max_{a=1,\dots,K} \widehat{\mu}_a(t)$ ayant la plus grande moyenne empirique et avec probabilité ε on tire un bras uniformément au hasard. Si ε est constant on a toujours un regret linéaire en T minoré par

$$\varepsilon \sum_{a=1}^{K} \Delta_a \frac{T}{K}\,.$$

Mais si l'on prend ε_t décroissant avec t, par exemple $\varepsilon_t = 6K/(d^2 t)$ où $0 < d < \min_{\Delta_a > 0} \Delta_a$, on peut montrer, voir Auer et al. [2002a], que le regret est au plus de l'ordre de $K \log(T)/d + o(T)$. Cependant cela nécessite de connaître à l'avance une borne inférieure sur les Δ_a.

Une deuxième solution consiste à construire une borne supérieure de confiance sur la moyenne de chaque bras avec un niveau de confiance soigneusement choisi, puis jouer le bras ayant la plus grande borne supérieure de confiance après une phase d'initialisation, plutôt que de comparer directement les moyennes empiriques. Cette méthode s'inspire du principe d'*optimisme en présence d'incertitude* voir Agrawal [1995], Burnetas and Katehakis [1996] et Munos et al. [2014]. Par exemple, pour $X_1, \dots, X_n$ i.i.d. selon une loi de Bernoulli de paramètre μ et $x < \mu$, l'inégalité de Hoeffding donne

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-n\,2(x-\mu)^2}.$$

Ce qui permet d'obtenir la borne supérieure de confiance de niveau δ : avec probabilité au moins $1-\delta$

$$\mu \leqslant \widehat{\mu}_n + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{1.1}$$

En prenant $\delta = 1/T$ on obtient la borne supérieure de confiance de l'algorithme UCB (Upper Confidence Bound), voir Auer et al. [2002a].

Algorithm 1: algorithme UCB

Initialisation: Tirer chaque bras de $\{1, \dots, K\}$ une fois.

Pour $t = K$ à $T-1$, **faire**

1. Calculer pour chaque bras a la borne supérieure de confiance

$$U_a^{\mathrm{UCB}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log(T)}{2N_a(t)}}.$$

2. Jouer $A_{t+1} \in \underset{a\in\{1,\cdots,K\}}{\arg\max}\, U_a^{\mathrm{UCB}}(t)$.

On a ajouté à la moyenne empirique un terme qui joue le rôle de bonus d'exploration pour les bras qui ont été peu tirés. Cet algorithme très simple appartient à la famille plus large des *politiques d'indice* (voir Gittins [1979]) où à chaque tour t, après éventuellement une phase d'initialisation, pour chaque bras a, l'agent construit un indice dépendant uniquement des récompenses issues de ce bras, ici la borne supérieure de confiance $U_a^{\mathrm{UCB}}(t)$, puis tire le bras ayant le plus grand indice. L'utilisation de borne supérieure de confiance comme indice a été étudiée, entre autres, par Lai and Robbins [1985]. Auer et al. [2002a] ont établi une borne *non-asymptotique* sur le regret du même type que celle de la Proposition 1.1.1. Initialement l'algorithme UCB a été conçu pour des récompenses bornées dans l'intervalle unité (voir la Section 1.2). L'algorithme 1 diffère de celui introduit par Auer et al. [2002a] en deux points. La constante devant le $\log(T)$ dans la définition de l'indice $U_a^{\mathrm{UCB}}(t)$ n'est pas la même, $1/2$ à la place de 2, cela pour faciliter la comparaison avec les indices des algorithmes qui vont suivre. Ce changement permet aussi d'obtenir une meilleure constante devant le terme en $\log(T)$ dans la borne de regret (au détriment du terme caché dans le $o\big(\log(T)\big)$). On a aussi choisi de présenter une version *non adaptative* en l'horizon, i.e., l'algorithme doit *connaître à l'avance l'horizon* T. Il existe des versions adaptatives en l'horizon, voir Auer et al. [2002a].

Proposition 1.1.1. *Pour l'algorithme UCB, pour tout bras a sous-optimal*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{2\Delta_a^2}\log(T) + o\big(\log(T)\big)\,.$$

Par conséquent

$$R_T \leqslant \sum_{a:\ \mu^\star>\mu_a} \frac{1}{2\Delta_a}\log(T) + o\big(\log(T)\big)\,.$$

Ainsi, pour l'algorithme UCB le regret croit non plus linéairement avec l'horizon T mais au plus *logarithmiquement*. On a même un résultat un peu plus fort : en espérance les bras sous-optimaux sont tirés au plus un nombre de fois proportionnel à $\log(T)$. Une question naturelle est alors de savoir si c'est le mieux que l'on puisse faire. Pour cela il faut établir une borne inférieure sur le regret.

1.1.2 Borne inférieure asymptotique sur le regret

Si aucune hypothèse n'est faite sur la stratégie suivie par l'agent le regret est trivialement minoré par zéro. En effet si l'agent décide de tirer uniquement le premier bras et que par chance ce dernier est optimal alors le regret est nul. Cependant cette stratégie peut s'avérer désastreuse. Si l'on permute ce bras avec un bras sous-optimal le regret est alors proportionnel à T. Il faut donc trouver un moyen d'éliminer ces stratégies triviales. Pour cela on va imposer le même type de garanties que l'on sait prouver pour l'algorithme UCB : en espérance les bras sous-optimaux sont tirés au plus $\log(T)$ fois. Moralement, on suppose que la stratégie fait toujours aussi bien que l'algorithme UCB et l'on va chercher à savoir si l'on peut faire mieux. Posons

$$\mathcal{D}_{ber} = \left\{ \underline{\nu} \ : \ \forall a \in \{1, \dots, K\}, \exists \mu_a \in [0,1] \text{ tel que } \nu_a = \mathrm{Ber}(\mu_a) \right\},$$

la collection des problèmes de bandit Bernoulli.

Définition 1.1.1. Une stratégie est *uniformément convergente* si pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, pour tout bras sous-optimal a, pour tout $0 < \alpha \leqslant 1$, elle satisfait $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] = o(T^\alpha)$.

On ajoute ici l'indice $\underline{\nu}$ à l'espérance pour préciser dans quel problème de bandit on se place. Il faut aussi définir une quantité qui permet de mesurer l'écart entre deux lois de Bernoulli de paramètre p et q dans $[0,1]$: la divergence de Kullback-Leibler,

$$\mathrm{kl}(p,q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

La borne inférieure de Lai and Robbins [1985] assure que le regret d'une stratégie uniformément convergente est de l'ordre de $\log(T)$.

Théorème 1.1.2. *(Borne inférieure de Lai et Robbins) Pour toute stratégie uniformément convergente, pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, pour tout bras sous-optimal a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log(T)} \geqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}.$$

On constate que la croissance en l'horizon T du regret est optimale pour l'algorithme UCB. Néanmoins la constante devant le $\log(T)$ n'est pas la même dans les deux bornes. En effet d'après l'inégalité de Pinsker

$$\mathrm{kl}(p,q) \geqslant 2(p-q)^2, \tag{1.2}$$

cette inégalité étant stricte pour $p \neq q$.

Définition 1.1.2. Une stratégie est *asymptotiquement optimale* sur $\mathcal{D}_{ber}$, si pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, tout bras sous optimal a,

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log(T)} \leqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}.$$

En particulier l'algorithme UCB n'est pas asymptotiquement optimal d'après (1.2). On va maintenant esquisser la preuve de la borne inférieure. En effet, cette dernière, de manière assez inhabituelle, éclaire assez bien ce qui se passe pour un horizon très grand. De plus, la plupart des algorithmes s'inspirent directement de cette dernière. Par exemple les index des différents algorithmes de type UCB sont construits en essayant d'optimiser cette borne avec les estimées courantes.

Esquisse de preuve du Théorème 1.1.2. Soit a un bras sous-optimal. Considérons un second de problème de bandit $\underline{\nu}'$ semblable au problème initial $\underline{\nu}$ où l'on a seulement changé la moyenne du bras a en $\mu' > \mu^\star$ de telle sorte que ce dernier soit optimal dans le nouveau problème. Ainsi on pose

$$\underline{\nu}' = \big(\mathrm{Ber}(\mu_1), \dots, \mathrm{Ber}(\mu'_a), \dots, \mathrm{Ber}(\mu_K)\big)\,.$$

Notons I_t l'information disponible par l'agent à l'instant t. Par exemple $I_t = (Y_1, \dots, Y_t)$ si sa stratégie est déterministe. Et notons $\mathbb{P}_{\underline{\nu}}^{I_{t+1}}$ respectivement $\mathbb{P}_{\underline{\nu}'}^{I_{t+1}}$ sa loi dans le problème $\underline{\nu}$ respectivement $\underline{\nu}'$. On va utiliser une conséquence très utile du principe de contraction de l'entropie.

Corollaire 1.1.3 (Contraction de l'entropie pour des espérances de variables aléatoires)**.** *Soit $\mathbb{P}$ et $\mathbb{Q}$ deux lois de probabilité définies sur le même espace mesurable $(\Omega, \mathcal{F})$, et soit X une variable aléatoire sur $(\Omega, \mathcal{F})$ à valeurs dans $[0,1]$. Posons $\mathbb{E}_{\mathbb{P}}[X]$ et $\mathbb{E}_{\mathbb{Q}}[X]$ l'espérance de X sous $\mathbb{P}$ et $\mathbb{Q}$ respectivement. Alors,*

$$\mathrm{kl}\big(\mathbb{E}_{\mathbb{P}}[X], \mathbb{E}_{\mathbb{Q}}[X]\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q})\,.$$

En conditionnant pour l'égalité et en utilisant le principe de contraction de l'entropie avec des espérances pour la première inégalité il vient

$$\begin{aligned}\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{kl}(\mu_a, \mu'_a) = \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_{t+1}}, \mathbb{P}_{\underline{\nu}'}^{I_{t+1}}\Big) &\geqslant \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}}\big[N_a(T)/T\big], \mathbb{E}_{\underline{\nu}'}\big[N_a(T)/T\big]\Big) \\ &\geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T}\right) \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]} - \log 2\,, \qquad (1.3)\end{aligned}$$

où l'on a utilisé $\mathrm{kl}(p,q) \geqslant (1-p)\log\big(1/(1-q)\big) - \log(2)$ pour la dernière inégalité. Il ne reste plus qu'à exploiter l'hypothèse de convergence uniforme. Puisque a est sous-optimal dans le problème $\underline{\nu}$ on a $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]/T \to 0$ et que a est optimal dans $\underline{\nu}'$, pour tout $0 < \alpha \leqslant 1$

$$\liminf_{T\to\infty} \frac{1}{\log T} \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]} \geqslant \liminf_{T\to\infty} \frac{1}{\log T} \log \frac{T}{T^\alpha} = (1-\alpha)\,.$$

En substituant cela dans (1.3), il vient, pour tout $\mu'_a > \mu^\star$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log T} \geqslant \frac{1}{\mathrm{kl}(\mu_a, \mu'_a)}\,.$$

□

1.1.3 kl-UCB un algorithme asymptotiquement optimal

La constante devant le $\log(T)$ dans la borne du regret de l'algorithme UCB provient de l'inégalité de Hoeffding utilisée pour construire les bornes supérieures de confiance. Pour prouver cette dernière on a majoré la variance d'une distribution à valeurs dans $[0,1]$ par $1/4$. Ce faisant on a aussi affaibli l'inégalité. La même majoration uniforme de la variance permet de prouver de manière duale l'inégalité de Pinsker. L'intuition est donc de choisir une divergence adaptée à la famille de distributions, i.e., garder la divergence de Kullback-Leibler dans l'inégalité de déviations. On obtient alors l'inégalité de Chernoff, pour $x < \mu$

$$\mathbb{P}(\widehat{\mu}_n < x) \leqslant e^{-n\,\mathrm{kl}(x,\mu)}, \tag{1.4}$$

avec $\widehat{\mu}_n = \sum_{k=1}^{n} X_k/n$ où $X_1, \ldots, X_n$ i.i.d. selon une loi de Bernoulli de paramètre μ. Cette inégalité est plus spécifique et donc plus forte que l'inégalité Hoeffding. On peut en déduire une inégalité de déviations pour la divergence kl : pour $u > 0$

$$\mathbb{P}\big(\widehat{\mu}_n < \mu \text{ et } \mathrm{kl}(\widehat{\mu}_n, \mu) > u\big) \leqslant e^{-nu}\,. \tag{1.5}$$

Puis inverser l'Inégalité (1.5) pour obtenir une nouvelle borne supérieure de confiance. Avec probabilité $1-\delta$, on a

$$\mu \leqslant \sup\big\{\mu' \geqslant \widehat{\mu}_n \;:\; n\mathrm{kl}(\widehat{\mu}_n, \mu') \leqslant \log(1/\delta)\big\}\,.$$

À la différence de (1.1) cette dernière n'est pas explicite. Toujours en prenant $\delta = 1/T$, on obtient l'indice de l'Algorithme 2 : kl-UCB de Cappé et al. [2013] et Burnetas and Katehakis [1996], voir aussi Lai and Robbins [1985].

Algorithm 2: Algorithme kl-UCB.

Initialisation: Tirer chaque bras de $\{1, .., K\}$ une fois.

Pour $t = K$ à $T-1$, **faire**

1. Calculer pour chaque bras a la quantité

$$U_a^{\mathrm{kl}}(t) = \sup\left\{\mu \geqslant \widehat{\mu}_a(t) \;:\; \mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \leqslant \frac{\log(T)}{N_a(t)}\right\}.$$

2. Jouer $A_{t+1} \in \underset{a\in\{1,..,K\}}{\arg\max}\, U_a^{\mathrm{kl}}(t)$.

On peut alors prouver que cet algorithme est asymptotiquement optimal, confer Garivier and Cappé [2011] et Cappé et al. [2013].

Proposition 1.1.4. *Pour l'algorithme kl-UCB, pour tout bras a sous-optimal*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}\log(T) + o\big(\log(T)\big)\,.$$

Il existe une multitude d'algorithmes asymptotiquement optimaux, dont une famille importante est les algorithmes d'inspiration bayésiennne. Le plus connu d'entre eux est l'algorithme de Thompson Sampling de Thompson [1933]. Il consiste à placer une loi a priori π_a^0 sur chacune des moyennes μ_a, typiquement une loi bêta $\pi_a^0 = \text{Beta}(\alpha, \beta)$. Puis à chaque tour t, l'agent tire un vecteur de moyennes selon la loi a posteriori courante Π^t et choisit le bras ayant la plus grande des moyennes, comme décrit dans l'Algorithme 3.

Algorithm 3: Algorithme de Thompson Sampling.

Paramètre: Une loi a priori sur les moyennes $\Pi^0 = (\pi_1^0, \dots, \pi_K^0)$.

Pour $t = 0$ à $T-1$, **faire**

1. **Pour** $a = 1$ à K, **faire**

   Tirer $\mu_a(t) \sim \pi_a^t$.

2. Jouer $A_{t+1} \in \underset{a \in \{1,..,K\}}{\arg\max}\, \mu_a(t)$, puis mettre à jour la loi a posteriori Π^{t+1}.

On peut alors montrer que si la loi a priori est $\Pi^0 = \big(\text{Beta}(1,1), \dots, \text{Beta}(1,1)\big)$ l'algorithme de Thompson Sampling est asympotiquement optimal, confer Korda et al. [2013]. Cet algorithme peut paraître a priori éloigné de l'algorithme kl-UCB, cependant il existe un troisième algorithme : Bayes-UCB, asymptotiquement optimal (voir Kaufmann et al. [2012]), qui fait le pont entre ces deux algorithmes. Il consiste, toujours après avoir placer un prior Π^0 sur les moyennes des bras, à tirer au tour t, le bras dont le quantile d'ordre $1 - 1/\big(t(\log t)^5\big)$ de la loi a posteriori courante est le plus grand, soit

$$A_{t+1} = \underset{a \in \{1,\dots,K\}}{\arg\max}\; Q\left(1 - \frac{1}{t(\log t)^5}; \Pi_a^t\right),$$

où $Q(\alpha; \pi)$ est le quantile d'ordre α de la distribution π. On peut alors montrer que les indices de kl-UCB est ceux de Bayes-UCB sont comparables aux perturbations près introduites par l'a priori, voir Kaufmann et al. [2012].

1.1.4 Les différents régimes du regret

Un défaut de l'analyse qui vient d'être présentée est d'être *asymptotique*. En effet, en se référant au Théorème 1.1.2 et à la Proposition 1.1.4 on pourrait penser que la croissance du regret devrait ressembler à $\log(T)$ multiplié par une certaine constante, cependant il apparaît clairement (voir Figure 1.1, gauche) que pour T petit ou modéré on n'obtient pas la forme logarithmique attendue. Même pour un horizon T grand les termes de second ordre continuent à jouer un rôle non négligeable en gardant le regret en dessous de la borne inférieure *asymptotique* de Lai et Robbins (voir Figure 1.1, droite).

En fait, on peut distinguer trois phases successives dans la croissance du regret : une phase initiale où les bras sont tirés de manière uniforme, une phase de transition lorsque le nombre d'observations devient suffisant pour détecter une différence entre les

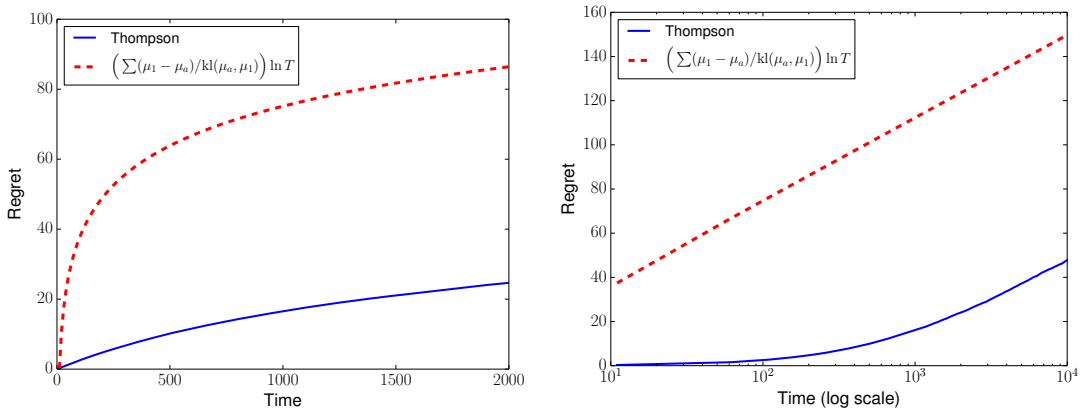

Figure 1.1: Regret moyen de l'algorithme Thompson [1933] Sampling (*bleu, plein*) pour un problème de bandit avec des lois de Bernoulli de paramètres $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$; les espérances sont approchées avec 500 expériences. Versus la borne inférieure *asymptotique* de Lai and Robbins [1985] (*rouge, pointillés*).

bras et une phase finale où l'on connaît les différentes moyennes des bras avec grande probabilité et que chaque nouveau tirage ne fait que confirmer l'identité du meilleur bras. La dernière phase est celle qui est décrite par la borne inférieure de Lai et Robbins (Théorème 1.1.2) où le regret croit de manière logarithmique. À l'opposé, lors de la phase initiale, la croissance du regret est linéaire en T. On peut aussi donner une borne inférieure qui décrit ce régime. Une stratégie est toujours meilleure que la stratégie uniforme sur $\mathcal{D}_{ber}$ si pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, pour tout bras optimal $a^\star$, pour tout $T \geqslant 1$

$$\mathbb{E}_{\underline{\nu}}\big[N_{a^\star}(T)\big] \geqslant \frac{T}{K}\,.$$

Proposition 1.1.5. *Pour toute stratégie meilleure que la stratégie uniforme, pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, pour tout bras a, pour tout $T \geqslant 1$,*

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geqslant \frac{T}{K}\Big(1 - \sqrt{2T\mathrm{kl}(\mu_a, \mu^\star)}\Big)\,.$$

En particulier,

$$\forall\, T \leqslant \frac{1}{8\mathrm{kl}(\mu_a, \mu^\star)}, \qquad \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geqslant \frac{T}{2K}\,.$$

1.1.5 Borne inférieure minimax

Un moyen de capturer la phase de transition entre le régime linéaire initial et le régime logarithmique est de s'intéresser au pire regret accumulé par une stratégie parmi tous les problèmes de bandit possibles. Autrement dit, on va étudier le risque minimax. Auer et al. [2002b] prouvent la borne inférieure suivante :

Théorème 1.1.6. *(Borne inférieure minimax) Pour toute stratégie*

$$\sup_{\underline{\nu}\in\mathcal{D}_{ber}} R_{T,\underline{\nu}} \geqslant \frac{1}{20}\min(\sqrt{KT}, T),$$

où le supremum est pris sur l'ensemble des problèmes de bandit de $\mathcal{D}_{ber}$.

Cette borne minimax tient à la fois dans le cadre des problèmes bandit stochastique et des problèmes de bandit non-stochastique introduit dans ce même papier Auer et al. [2002b]. À noter que pour un horizon T fixé le regret minimax est atteint pour un problème de bandit où les écarts aux meilleurs bras sont de l'ordre de $\sqrt{K/T}$. Dès lors il est naturel de définir aussi l'optimalité pour le risque minimax.

Définition 1.1.3. Une stratégie est minimax optimale sur $\mathcal{D}_{ber}$, s'il existe une constante C telle que pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{ber}$, pour tout T,

$$R_{T,\underline{\nu}} \leqslant C\sqrt{KT}\,.$$

On peut noter une différence fondamentale entre ces deux notions d'optimalité. D'un coté l'optimalité asymptotique dépend profondément du problème considéré via la constante mais est, comme son nom l'indique, asymptotique. De l'autre coté, l'optimalité minimax tient pour un horizon T fixé mais est indépendante du problème considéré. L'enjeu est alors de trouver un algorithme *simultanément* asymptotiquement et minimax optimal. C'est une des contributions de cette thèse.

1.1.6 kl-UCB^{++} un algorithme minimax et asymptotiquement optimal

On peut montrer la borne supérieure suivante sur le regret pour l'algorithme UCB :

$$R_T \leqslant C'\sqrt{KT\log(T)}\,.$$

pour une certaine constante C'. Il y a donc un facteur $\log(T)$ en trop pour qu'il soit minimax optimal. Pour remédier à ce problème il suffit de changer un peu l'exploration. On obtient alors l'algorithme MOSS (Minimax Optimal Strategy in the Stochastic case) introduit par Audibert and Bubeck [2009].

Algorithm 4: Algorithme MOSS.

Initialisation: Tirer chaque bras de $\{1, .., K\}$ une fois.

Pour $t = K$ à $T-1$, **faire**

1. Calculer pour chaque bras a la quantité

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log_+\Big(T/\big(KN_a(t)\big)\Big)}{2N_a(t)}}.$$

2. Jouer $A_{t+1} \in \underset{a\in\{1,..,K\}}{\arg\max}\, U_a^{\mathrm{M}}(t)$.

Où l'on note $\log_+(x) = \max\big(\log(x), 0\big)$. À noter que l'on utilise une constante légèrement différente de celle du papier originel [Audibert and Bubeck, 2009] pour rester cohérent avec celle choisie dans l'algorithme UCB. La division par $N_a(t)$ dans le log est une réminiscence de la fonction d'exploration utilisée dans Lai [1987]. Il s'agit ici d'annuler le bonus d'exploration lorsqu'un bras a été tiré plus de T/K fois. En effet puisque dans le pire des cas les bras sous-optimaux sont à une distance $\sqrt{K/T}$ des bras optimaux, les moyennes empiriques seront séparées lorsque la taille de l'échantillon sera de l'ordre de T/K. On peut alors se contenter de prendre les moyennes empiriques dans cette situation. Audibert and Bubeck [2009] montre que l'algorithme MOSS est minimax optimal.

Proposition 1.1.7. *Pour l'algorithme MOSS*

$$R_T \leqslant 17\sqrt{KT} + K\,.$$

L'algorithme MOSS doit aussi connaître à l'avance l'horizon, mais on peut facilement le rendre adaptatif en remplaçant T par t dans la fonction d'exploration, soit l'indice

$$U_a^{\text{M-A}}(t) = \widehat{\mu}_{a,N_a(t)} + \sqrt{\frac{\log_+\Big(t/\big(KN_a(t)\big)\Big)}{2N_a(t)}}.$$

On peut montrer le même type de garanties que celles de la Proposition 1.1.7 pour cette version adaptative, voir la Section 4.5.3. Cependant l'algorithme MOSS partage le même défaut concernant le choix de la divergence que l'algorithme UCB. Il est donc difficile d'espérer prouver l'optimalité asymptotique de ce dernier. Mais on peut combiner cette modification du taux d'exploration avec l'algorithme kl-UCB pour obtenir l'algorithme kl-UCB^{++} ayant pour indice

$$U_a^{\text{kl}++}(t) = \sup\left\{\mu \geqslant \widehat{\mu}_a(t) \;:\; \mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \leqslant \frac{\log_+\Big(T/\big(KN_a(t)\big)\Big)}{N_a(t)}\right\}. \tag{1.6}$$

Ce dernier est une légère variation de l'algorithme kl-UCB$^+$ introduit par Garivier and Cappé [2011] voir aussi Lai [1987] où l'on a seulement ajouté un facteur K dans la fonction d'exploration :

$$U_a^{\text{kl}+}(t) = \sup\left\{\mu \geqslant \widehat{\mu}_a(t) \;:\; \mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \leqslant \frac{\log_+\big(T/N_a(t)\big)}{N_a(t)}\right\}. \tag{1.7}$$

On peut alors montrer qu'il est *simultanément* asymptotiquement optimal et minimax optimal.

Proposition 1.1.8. *Pour l'algorithme kl-UCB^{++}*

$$R_T \leqslant 17\sqrt{KT} + K\,. \tag{1.8}$$

De plus, pour tout bras a sous-optimal,

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)} + o\big(\log(T)\big)\,. \tag{1.9}$$

Un résultat similaire est prouvé dans Garivier et al. [2016] pour le cas particulier d'un problème de bandit avec deux bras gaussiens. De même l'algorithme AdaUCB (adaptive UCB) de Lattimore [2018] est à la fois minimax optimal, asymptotiquement optimal et vérifie une version non asymptotique de (1.9) (une troisième notion d'optimalité) pour des problèmes de bandit Gaussien. Voir aussi Bubeck and Slivkins [2012] pour le même type de garanties à la fois dans le cadre stochastique et non-stochastique.

1.1.7 Raffinements du terme de second ordre

Après avoir obtenu le premier terme du développement asymptotique du regret, on peut se demander ce qui se passe avec le terme suivant. Cela est d'autant plus intéressant que l'on a vu dans la Section 1.1.4 qu'il avait un effet non négligeable sur le comportement du regret même pour des horizons relativement grands. Tout comme il a été nécessaire de faire une hypothèse sur la stratégie pour obtenir le premier terme il faut faire de même pour le second. On a d'ailleurs besoin d'une hypothèse encore plus forte. Par exemple, il existe un constante C qui dépend de la stratégie telle que pour tout problème de bandit $\underline{\nu}$, tout bras sous-optimal a :

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \leqslant \frac{C\log(T)}{\Delta_a^2}\,. \tag{1.10}$$

Cette hypothèse reste malgré tout naturelle au regard des bornes supérieures obtenues précédemment, cf. Proposition 1.1.1. On peut alors montrer la borne inférieure suivante, voir Théorème 2.4.3.

Théorème 1.1.9. *Pour toute stratégie vérifiant* (1.10)*, pour tout problème de bandit* $\underline{\nu}$*,*

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geqslant \frac{\log T}{\mathrm{kl}(\mu_a,\mu^\star)} - O\big(\mathrm{loglog}(T)\big)\,. \tag{1.11}$$

Il se trouve que l'astuce consistant à diviser par $N_a(t)$ dans la fonction d'exploration permet aussi d'obtenir le bon second ordre de grandeur dans le développent asymptotique du regret. En effet pour l'algorithme kl-UCB$^+$ on peut montrer le théorème suivant.

Proposition 1.1.10. *Pour l'algorithme* kl-*UCB*$^+$*, pour tout bras* a *sous-optimal,*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T) - \mathrm{loglog}(T)}{\mathrm{kl}(\mu_a,\mu^\star)} + O(1)\,. \tag{1.12}$$

On comprend alors pourquoi dans la Figure 1.1 le regret reste en dessous de la borne asymptotique de Lai et Robbins. Une borne similaire tient aussi pour l'algorithme kl-UCB^{++} puisque le facteur K additionnel n'intervient pas dans le régime asymptotique. La première borne sur le regret exhibant le bon second ordre a été prouvée par Honda and Takemura [2015] pour l'algorithme IMED. Le principe de la preuve est le suivant :

pour un bras a sous-optimal, on décompose l'espérance de la façon suivante :

$$\mathbb{E}\big[N_a(T)\big] \leqslant 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(U_a(t) \leqslant \mu^\star - \delta,\ A_{t+1} = a\big)}_{\mathrm{A}} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(\mu^\star - \delta < U_a(t),\ A_{t+1} = a\big)}_{\mathrm{B}}, \tag{1.13}$$

avec $\delta > 0$ un paramètre à fixer. Pour le terme A la nouveauté est d'utiliser les déviations de l'indice $U_a(t)$ au lieu de directement minorer $U_a(t)$ par $U_{a^\star}(t)$. En effet, si $N_a(t)$ est grand, $U_a(t)$ est de l'ordre de μ_a et donc la probabilité apparaissant dans le terme A est faible. On peut alors se contenter de traiter le cas $N_a(t) \lesssim \log(T)$. Ce qui permet de choisir $\delta \sim 1/\log(T)$ le bon ordre de grandeur pour obtenir le second terme en $-\operatorname{loglog}(T)$. Pour le terme B, on suit les mêmes arguments introduit par Honda and Takemura [2015, Lemme 18]. L'idée est d'exploiter le fait que c'est le même processus qui intervient dans chacune des probabilités du terme B et non de les majorer séparément.

Donnons maintenant une intuition sur la fonction d'exploration via la preuve de la borne inférieure de Lai et Robbins (Théorème 1.1.2). On se place dans le problème $\underline{\nu}$ et on considère un problème alternatif $\underline{\nu}'$ identique au problème initial excepté que l'on a déplacé la moyenne du bras a au-dessus de $\mu^\star < \mu'_a$. Le bras a est donc l'unique bras optimal dans $\underline{\nu}'$. Réécrivons l'Inégalité (1.3) de la preuve

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{kl}(\mu_a, \mu'_a) \geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T}\right) \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]} - \log 2\,.$$

Si l'on suppose que les estimées sont égales ou très proches des vraies valeurs :

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \approx N_a(T) \quad \left(\approx \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^\star)}\right)$$
$$\mu_a \approx \widehat{\mu}_a(t)\,,$$

on obtient pour tout $\mu'_a > \widehat{\mu}_a(t)$,

$$N_a(T)\,\mathrm{kl}(\widehat{\mu}_a(t), \mu'_a) \gtrsim \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]}\,. \tag{1.14}$$

C'est exactement ce type d'inégalité que l'on optimise pour construire la borne supérieure de confiance. Par exemple pour l'algorithme kl-UCB^{++} :

$$\begin{aligned} U_a^{\mathrm{kl}++}(t) &= \sup\left\{\mu \geqslant \widehat{\mu}_a(t)\ :\ N_a(t)\,\mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \leqslant \log_+ \frac{T}{KN_a(t)}\right\} \\ &= \inf\left\{\mu \geqslant \widehat{\mu}_a(t)\ :\ N_a(t)\,\mathrm{kl}\big(\widehat{\mu}_a(t), \mu\big) \geqslant \log_+ \frac{T}{KN_a(t)}\right\}, \end{aligned}$$

à noter que les indices de type UCB ont été introduits sous la seconde forme, confer Lai [1987]. Pour parfaire le parallèle, il reste à identifier la fonction d'exploration avec le

second terme de (1.14). Idéalement on souhaiterait utiliser ce dernier, mais on n'a pas accès à une estimée naturelle de $\mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]$. Cependant on a

$$T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big] = \sum_{b \neq a} \mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big] \,.$$

Puisque dans $\underline{\nu}'$ seul a est optimal et que dans $\underline{\nu}$ le bras a est sous-optimal, pour tout $b \neq a$ on approche brutalement $\mathbb{E}_{\underline{\nu}'}\big[N_b(T)\big] \approx N_a(T)$, d'où

$$T - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big] \approx K N_a(t) \,,$$

ce qui permet de retrouver l'indice de kl-UCB^{++}. On peut aussi se référer à Lattimore [2018] pour une autre interprétation, moins asymptotique, de la fonction d'exploration grâce à une borne inférieure.

1.1.8 Perspectives

Plusieurs pistes restent à explorer. Une d'entre elles pourrait être d'obtenir une borne supérieure sur le regret qui se spécifierait en la borne minimax ou asymptotique selon le régime considéré. Cela permettrai de mieux comprendre le comportement du regret entre ces deux régimes. Une première étape vers ce type de résultat pourrait être d'adapter la troisième notion d'optimalité proposée par Lattimore [2018] pour les problèmes de bandit Bernoulli et de montrer qu'un algorithme de type kl-UCB (à une modification près de la fonction d'exploration) atteint cette dernière.

Une autre piste naturelle serait de raffiner l'analyse présentée en Section 1.1.7 afin d'obtenir la constante optimale devant le terme de second ordre. Pour cela il faudrait aussi trouver la bonne fonction d'exploration à utiliser dans l'index de l'algorithme kl-UCB. Un moyen d'y parvenir serait de pousser l'analogie avec la borne inférieure de Lai et Robbins à l'ordre deux, plus précisément la borne inférieure non asymptotique du Théorème 2.4.3.

1.2 Problème de bandit non-paramétrique

On considère maintenant une extension non-paramétrique du cadre précédent. On supposera seulement que les récompenses sont bornées dans l'intervalle unité. Le problème de bandit $\underline{\nu}$ sera donc une collection de K bras chacun associé à une distribution ν_a à support dans $[0,1]$, de moyenne $\mu_a := E(\nu_a)$. De la même façon que précédemment on pose $\mathcal{D}_{bor}$ la collection des problèmes de bandit borné. Ce n'est qu'un cadre parmi tant d'autres, voir par exemple Lattimore [2017] et les références citées. Il a été étudié par exemple par Auer et al. [2002a] et Honda and Takemura [2010]. On rappelle que la divergence de Kullback-Leibler entre deux distributions $\mathbb{P}$ et $\mathbb{Q}$ est définie par

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \begin{cases} \displaystyle\int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{P} & \text{si } \mathbb{P} \ll \mathbb{Q}; \\ +\infty & \text{sinon}\,. \end{cases}$$

1.2.1 Borne inférieure asymptotique

Burnetas and Katehakis [1996] généralise le Théorème 1.1.2 à un cadre non-paramétrique qui englobe celui des problèmes de bandit borné.

Théorème 1.2.1. *Pour toute stratégie uniformément convergente sur $\mathcal{D}_{bor}$, pour tout problème de bandit $\underline{\nu} \in \mathcal{D}_{bor}$, pour tout bras sous-optimal a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log(T)} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

L'ordre de grandeur en l'horizon reste inchangé mais la constante caractérisant la complexité du problème est différente. Cette nouvelle quantité, voir Figure 1.2, est définie pour $\nu \in \mathcal{P}[0,1]$ et $\mu \in [0,1[$ par

$$\mathcal{K}_{\inf}(\nu, \mu) := \inf\big\{\mathrm{KL}(\nu, \nu') :\ \nu' \in \mathcal{P}[0,1],\ E(\nu') > \mu\big\}\,. \tag{1.15}$$

C'est l'infimum des divergences de Kullback-Leibler entre la distribution ν et un élément du demi-espace des distributions ayant une moyenne plus grande que μ. En fait, l'infimum est atteint pour une certaine distribution $\nu_\mu^\star$ de moyenne μ. On peut interpréter cette dernière comme une projection de ν sur le demi-espace défini ci-dessus pour la divergence de Kullback-Leibler. Il est intéressant de remarquer que ce n'est pas la projection habituellement considérée puisque ici les arguments sont inversés (voir Csiszár and Matus [2003]). De la même façon que précédemment on dira qu'un algorithme est asymptotiquement optimal si

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log(T)} \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

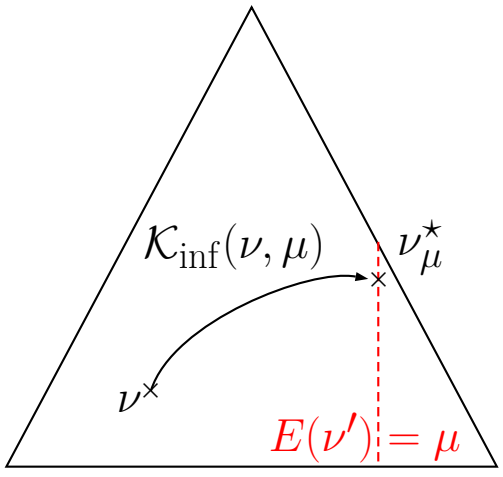

Figure 1.2: Construction de $\mathcal{K}_{\inf}(\nu, \mu)$

En comparant cette nouvelle borne inférieure et la borne sur le regret de la Proposition 1.1.4, il apparaît que l'algorithme kl-UCB n'est pas asymptotiquement optimal pour ce nouveau cadre. À noter que l'on peut utiliser cet algorithme avec des récompenses bornées puisque ce dernier se sert uniquement des moyennes empiriques pour construire les indices des différents bras (voir Garivier and Cappé [2011]). En effet, on dispose de l'inégalité suivante par contraction de l'entropie :

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \geqslant \mathrm{kl}\big(E(\nu_a),\, \mu^\star\big)\,,$$

avec égalité si ν_a est une loi de Bernoulli. Que l'on peut réécrire

$$\underbrace{\inf\Big\{\mathrm{KL}(\nu, \nu') :\ E(\nu') > \mu\Big\}}_{\mathcal{K}_{\inf}(\nu,\mu)} \geqslant \underbrace{\inf\Big\{\mathrm{KL}(\nu'', \nu') :\ E(\nu') > \mu,\, E(\nu'') = E(\nu)\Big\}}_{\mathrm{kl}\left(E(\nu),\,\mu\right)}\,.$$

L'intuition est la suivante : kl-UCB est sous-optimal car la seule information qu'il extrait des observations est la moyenne empirique. Pour obtenir un algorithme asymptotiquement optimal il faudra utiliser toute l'information disponible : *la mesure empirique*.

1.2.2 Algorithme KL-UCB

L'idée pour ce nouvel algorithme est donc de remplacer la moyenne empirique par la mesure empirique et la divergence entre deux lois de Bernoulli par la divergence de Kullback-Leibler générale dans l'indice de l'algorithme kl-UCB. En notant $\widehat{\nu}(t) = \big(1/N_a(t)\big)\sum_{s=1}^{t} \mathbb{1}_{\{A_s=a\}}\delta_{Y_s}$ la mesure empirique on définit le nouvel indice :

$$\begin{aligned} U_a^{\mathrm{KL}}(t) :&= \sup\left\{E(\nu') \geqslant E\big(\widehat{\nu}_a(t)\big) :\ \nu' \in \mathcal{P}[0,1],\ \ \mathrm{KL}\big(\widehat{\nu}_a(t),\, \nu'\big) \leqslant \frac{\log(T)}{N_a(t)}\right\} \\ &= \sup\left\{\mu' :\ \mu' \in [0,1],\, \mu' \geqslant \widehat{\mu}_a(t),\ \ \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t),\, \mu'\big) \leqslant \frac{\log(T)}{N_a(t)}\right\}, \end{aligned} \tag{1.16}$$

où l'on a utilisé la définition et quelques propriétés de régularité de $\mathcal{K}_{\inf}\big(\widehat{\nu}_a(t),\, \mu'\big)$ pour la seconde égalité. Cette borne supérieure de confiance est très similaire à celle que l'on pourrait obtenir en utilisant la méthode de la vraisemblance empirique introduite par Owen [1990]. À la différence près, non négligeable ici, qu'avec la vraisemblance empirique on se restreint aux distributions ν' à support dans l'enveloppe convexe du support de $\widehat{\nu}_a(t)$. L'utilisation de la vraisemblance empirique dans le cadre des problèmes de bandit stochastique a été introduite par Honda and Takemura [2010].

On obtient alors un nouvel algorithme similaire à l'Algorithme 2 où seul l'indice a été modifié. Ce dernier a été introduit par Maillard et al. [2011] et Cappé et al. [2013] pour des distributions à support borné expérimentalement et traité théoriquement seulement pour les distributions à support fini.

Algorithm 5: Algorithme KL-UCB.

Initialisation: Tirer chaque bras de $\{1, .., K\}$ une fois.

Pour $t = K$ à $T - 1$, **faire**

1. Calculer pour chaque bras a la quantité

$$U_a^{\mathrm{KL}}(t) = \sup\left\{\mu \in [0,1] \ : \ \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{\log(T)}{N_a(t)}\right\}.$$

2. Jouer $A_{t+1} \in \underset{a\in\{1,..,K\}}{\arg\max}\, U_a^{\mathrm{KL}}(t)$.

On peut alors montrer que cet algorithme est asymptotiquement optimal en généralisant les résultats de Cappé et al. [2013] aux distributions à support dans l'intervalle unité.

Proposition 1.2.2. *Pour l'algorithme KL-UCB, pour tout bras a sous-optimal*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\log(T) + o\big(\log(T)\big)\,.$$

Pour ce cadre il existe beaucoup moins d'algorithmes asymptotiquement optimaux, on peut par exemple citer l'algorithme DMED de Honda and Takemura [2010] ou IMED de Honda and Takemura [2015] et son extension aux distributions à support semi-borné. On pourrait aussi penser à une version non-paramétrique de l'algorithme Thompson Sampling où l'on aurait remplacé les a priori sur les moyennes des bras par une loi a priori sur les bras. Un choix naturel pourrait être de prendre comme a priori un processus de Dirichlet (voir Ferguson [1973]). Savoir si cette extension est aussi asymptotiquement optimale est une question ouverte.

Tout comme l'algorithme kl-UCB est associé à une inégalité de déviations (1.5) pour la divergence kl, l'indice (1.16) de l'algorithme KL-UCB est associé à une inégalité de déviation pour $\mathcal{K}_{\inf}$. En effet pour $X_1, \ldots, X_n$ i.i.d. selon ν et en notant $\widehat{\nu}_n = (1/n)\sum_{k=1}^n \delta_{X_k}$ on a

$$\mathbb{P}\Big(\mathcal{K}_{\inf}\big(\widehat{\nu}_n, E(\nu)\big) > u\Big) \leqslant e(2n+1)e^{-nu}\,. \tag{1.17}$$

Comme le montre la Figure 1.3 on cherche à majorer la probabilité que la mesure empirique $\widehat{\nu}_n$ appartienne à l'ensemble $\big\{\nu' : \ \mathcal{K}_{\inf}\big(\nu', E(\nu)\big) \geqslant u\big\}$. La difficulté majeure est que cet ensemble n'est pas convexe. On ne peut donc pas utiliser directement une inégalité du type inégalité de Sanov [Csiszár, 1984]. Une question ouverte est de savoir s'il est possible *de supprimer le facteur n supplémentaire* devant l'exponentielle. Pour le moment, le mieux que l'on puisse faire est remplacer ce facteur n par $\sqrt{n}$. Une piste serait d'étudier finement le comportement de la frontière de l'ensemble $\big\{\nu' : \ \mathcal{K}_{\inf}\big(\nu', E(\nu)\big) \geqslant u\big\}$ au voisinage du point de cet ensemble le plus proche de ν au sens de la divergence de Kullback-Leibler. Cela en s'inspirant de ce qui est fait dans Iltis [1995].

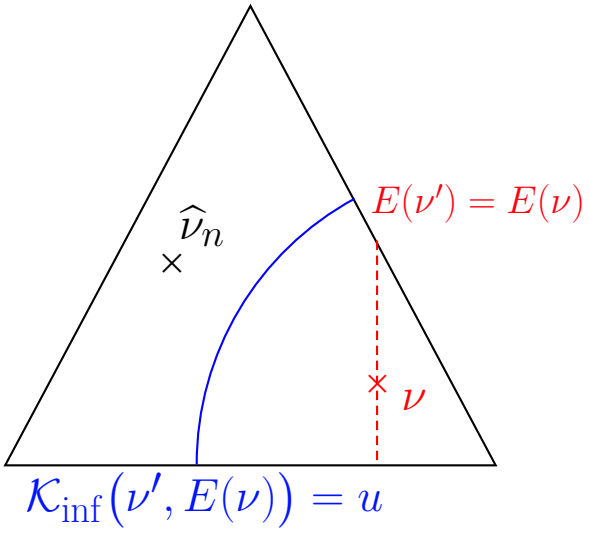

Figure 1.3: Inégalité de déviations pour $\mathcal{K}_{\inf}$.

À l'instar de la divergence de Kullback-Leibler, $\mathcal{K}_{\inf}$ possède aussi une formulation variationnelle :

$$\mathcal{K}_{\inf}(\nu, \mu) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}_\nu \left[\log\left(1 - \lambda \frac{X - \mu}{1 - \mu}\right) \right]. \tag{1.18}$$

Il est naturel de pouvoir réécrire $\mathcal{K}_{\inf}$ qui est par définition un infimum en maximum. En effet, d'après la formulation variationnelle de la divergence de Kullback-Leibler (voir Section 4.7.4), cette dernière peut s'exprimer comme un supremum et donc $\mathcal{K}_{\inf}$ comme un inf sup, que l'on peut permuter sous certaines conditions (cf. lemme de Sion) en sup inf. Cette formulation variationnelle est un élément clé pour prouver l'Inégalité (1.17). En effet, grâce à celle-ci on peut approcher l'ensemble $\big\{\nu' : \mathcal{K}_{\inf}\big(\nu', E(\nu)\big) \geqslant u\big\}$ par une union de demi-plans. Ensuite il suffit d'appliquer la borne de l'union puis l'inégalité de Sanov par exemple (puisque l'on s'est ramené à des convexes). Ce qui explique au passage le facteur n additionnel. En fait on peut se contenter de l'inégalité de Markov, voir la preuve de la Proposition 4.5.6. La formulation variationnelle, telle que présentée ici, a été originellement prouvée dans Honda and Takemura [2010]. Mais cette formulation était déjà connue dans la littérature de la vraisemblance empirique Harari-Kermadec [2006] ou encore en optimisation Borwein and Lewis [1991] et dans l'études des grandes déviations Pandit and Meyn [2006].

1.2.3 Optimalité minimax

Il est naturel, toujours dans l'optique de mimer le cadre paramétrique, de chercher un algorithme à la fois asymptotiquement optimal et minimax optimal. On peut par exemple transposer l'algorithme kl-UCB^{++} (1.6) au cadre non-paramétrique en définissant

l'indice de l'algorithme KL-UCB^{++}

$$U_a^{\mathrm{KL}++}(t) := \sup\left\{ \mu \in [0,1] \ : \ \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)} \log_+\left(\frac{T}{KN_a(t)}\right) \right\}. \tag{1.19}$$

Cependant l'inégalité de déviations pour $\mathcal{K}_{\inf}$ (1.17) est trop faible pour pouvoir montrer les mêmes résultats avec cet algorithme. On est alors contraint de définir un nouvel algorithme KL-UCB-switch dont l'indice est un hybride entre celui de KL-UCB^{++} et celui de MOSS :

$$U_a^{\mathrm{KL-S}}(t) = \begin{cases} U_a^{\mathrm{KL}++}(t) & \text{si } N_a(t) \leqslant f(T,K) \\ U_a^{\mathrm{M}}(t) & \text{si } N_a(t) > f(T,K) \end{cases},$$

où $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$. Autrement dit, lorsque le bras est tiré peu de fois son indice est celui de KL-UCB^{++} tandis que lorsqu'il est tiré un grand nombre de fois c'est celui de MOSS. On peut alors montrer que ce nouvel algorithme est à la fois minimax et asymptotiquement optimal (voir Théorème 4.2.1 et Théorème 4.2.2).

Proposition 1.2.3. *Pour l'algorithme KL-UCB-switch*

$$R_T \leqslant (K-1) + 25\sqrt{KT}\,.$$

De plus pour tout bras a sous-optimal

$$\mathbb{E}[N_a(T)] \leqslant \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o\big(\log(T)\big)\,.$$

Cette interpolation entre les deux indices est sans doute qu'un artifice technique et le même type de théorème doit aussi être valable pour l'algorithme KL-UCB^{++}. Cela reste cependant une question ouverte. On peut aussi montrer que MOSS et kl-UCB^{++} sont minimax optimal pour des récompenses bornées mais a priori non asymptotiquement optimal.

En s'appuyant sur les mêmes méthodes que celles de la Section 1.1.7 on peut aussi obtenir l'ordre de grandeur optimal pour terme de second ordre (voir Théorème 4.2.3).

Théorème 1.2.4. *Pour l'algorithme KL-UCB-switch*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\log(T) - \operatorname{loglog}(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + O(1)\,.$$

On peut rendre l'algorithme KL-UCB-switch adaptatif en l'horizon en modifiant légèrement la fonction d'exploration et la condition pour permuter entre l'indice de KL-UCB^{++} et celui de MOSS. Il est possible de montrer que l'algorithme ainsi obtenu est aussi simultanément minimax et asymptotiquement optimal, voir la Section 4.2.3. Cependant c'est une question ouverte de savoir si l'on peut aussi obtenir le bon second ordre comme dans le Théorème 1.2.4 pour cet algorithme.

1.2.4 Perspectives

La principale question soulevée précédemment est de savoir si l'on peut prouver l'optimalité minimax de l'algorithme KL-UCB^{++}. Une première étape non triviale serait de montrer une inégalité de déviation pour $\mathcal{K}_{\inf}$ du même type que (1.17) sans le facteur n additionnel. Une piste pourrait être de s'inspirer des travaux de Iltis [1995], comme il est dit en Section 1.2.2.

Dans une autre direction il serait intéressant de prouver l'optimalité asymptotique de l'algorithme de type Thompson Sampling brièvement décrit en Section 1.2.2.

Enfin on pourrait essayer de reproduire les résultats obtenus dans d'autres cadres non-paramétriques. Par exemple on pourrait considérer les problèmes de bandit semi-bornés à l'instar de Honda and Takemura [2015]. L'optimalité minimax n'a plus vraiment de sens, du moins sans hypothèse supplémentaire, dans ce nouveau cadre, mais les résultats concernant l'optimalité asymptotique devraient pouvoir se généraliser.

1.3 Bandit à seuil.

On va maintenant s'intéresser à *l'identification du meilleur bras*. En effet plutôt que de chercher à minimiser le regret, on peut seulement vouloir trouver un bras optimal en explorant le plus efficacement possible les distributions associées à chacun des bras. Pour ce problème deux approches parallèles coexistent. Soit on fixe un nombre T de tirages à l'avance et l'on essaye de prédire le meilleur bras avec la plus grande probabilité possible après ces T tirages. C'est le problème d'identification du meilleur bras à *budget fixé* introduit par Bubeck et al. [2012] et Audibert and Bubeck [2010], voir aussi Carpentier and Locatelli [2016] pour une analyse de la complexité de ce problème. Soit on impose de trouver le meilleur bras avec probabilité au moins $1-\delta$ et l'on essaye de minimiser le nombre de tirages que l'on doit faire pour y parvenir. C'est le problème d'identification du meilleur bras à *niveau de confiance fixé* introduit par Even-Dar et al. [2002] et Mannor and Tsitsiklis [2004b], voir aussi Kaufmann et al. [2016] pour une première analyse asymptotique (lorsque le niveau de confiance tend vers 0). La complexité asymptotique de ce problème a été établie par Garivier and Kaufmann [2016] dans le cadre fréquentiste et par Russo [2016] dans un cadre bayésien. Par la suite on suivra la seconde approche.

Plutôt qu'identifier le bras avec la plus grande moyenne on va chercher à trouver le bras ayant *la moyenne la plus proche possible d'un certain seuil* connu par l'agent. De plus on va supposer que *la moyenne des bras est une fonction croissante de l'indice*. C'est le problème de *bandit à seuil*. Voir Locatelli et al. [2016] et les références citées pour une introduction à ce type de problèmes. Ce dernier a une motivation pratique: la phase I des essais cliniques. En effet, elle consiste à déterminer la dose maximale admissible d'un médicament. C'est à dire la quantité maximum de ce même médicament que l'on peut administrer à un patient avant que les effets secondaires ne deviennent insupportables ou dangereux. Typiquement un seuil de tolérance est choisi et le but de l'essai clinique est d'*identifier rapidement* le dosage qui induit la toxicité la plus proche de ce seuil, la toxicité étant une *fonction croissante du dosage*. Habituellement l'essai est mené en

augmentant graduellement le dosage suivant le traditionnel plan d'expérience "3+3" voir Le Tourneau et al. [2009] et Genovese et al. [2013].

1.3.1 Cadre

On considère un problème de bandit Gaussien à $K \geqslant 2$ bras $\underline{\nu} = \big(\mathcal{N}(\mu_1, 1), \dots, \mathcal{N}(\mu_K, 1)\big)$ que l'on identifiera sans ambiguïté à son vecteur de moyennes $\boldsymbol{\mu} = \big(\mu_1, \dots, \mu_K\big)$. On note $\mathbb{P}_{\boldsymbol{\mu}}$ et $\mathbb{E}_{\boldsymbol{\mu}}$ respectivement la probabilité et l'espérance dans le problème de bandit $\boldsymbol{\mu}$. Un seuil $S \in \mathbb{R}$ est fixé, et on notera $a^\star_{\boldsymbol{\mu}} \in \arg\min_{1 \leqslant a \leqslant K} |\mu_a - S|$ un bras optimal pour ce seuil.

Soit $\mathcal{M}$ l'ensemble des problèmes de bandit Gaussien ayant un unique bras optimal et soit $\mathcal{I} = \{\boldsymbol{\mu} \in \mathcal{M} : \mu_1 < ... < \mu_K\}$ le sous-ensemble des problèmes ayant des moyennes croissantes. Ce dernier ensemble incorpore l'hypothèse selon laquelle la toxicité est une fonction croissante du dosage.

Définition d'un algorithme δ-correct. On fixe $\delta \in (0, 1)$ un niveau de confiance et un problème de bandit $\boldsymbol{\mu} \in \mathcal{M}$ ou $\mathcal{I}$. Le jeu se déroule de la manière suivante : à chaque tour $t \in \mathbb{N}^*$ l'agent choisit un bras $A_t \in \{1, \dots, K\}$ et reçoit une récompense $Y_t \sim \mathcal{N}(\mu_{A_t}, 1)$ conditionnellement indépendante du passé. Soit $\mathcal{F}_t = \sigma(A_1, Y_1, \dots, A_t, Y_t)$ l'information à disposition de l'agent à l'instant t. Son objectif est alors d'identifier le bras optimal $a^\star_{\boldsymbol{\mu}}$ tout en minimisant le nombre de tirage τ_δ. Pour cela l'agent doit définir :

- une **règle d'échantillonnage** $(A_t)_{t \geqslant 1}$, où A_t est $\mathcal{F}_{t-1}$-mesurable,
- un **critère d'arrêt** τ_δ, qui est un temps d'arrêt pour la filtration $(\mathcal{F}_t)_{t \geqslant 1}$,
- une **règle de décision** $\widehat{a}_{\tau_\delta}$ $\mathcal{F}_{\tau_\delta}$-mesurable.

Quelque soit le cadre $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$ (problème non-monotone ou croissant), un algorithme est dit δ-correct sur $\mathcal{S}$ si pour tout $\boldsymbol{\mu} \in \mathcal{S}$ on a $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < +\infty) = 1$ et $\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta} \neq a^\star_{\boldsymbol{\mu}}) \leqslant \delta$. Étant donné un algorithme δ-correct on juge son efficacité à travers $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$.

1.3.2 Borne inférieure

Pour $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$, on définit l'ensemble des *alternatives* au problème de bandit $\boldsymbol{\mu} \in \mathcal{M}$ par

$$\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S}) := \{\boldsymbol{\lambda} \in \mathcal{S} : a^\star_{\boldsymbol{\lambda}} \neq a^\star_{\boldsymbol{\mu}}\}\,, \tag{1.20}$$

et Σ_K le simplexe dimension de $K-1$. De la même façon que l'on peut minorer le nombre de fois qu'un bras sous-optimal est tiré avec la borne inférieure de Lai et Robbins, on peut minorer le nombre moyen de tirages nécessaires pour atteindre un niveau de confiance fixé. Ce sont d'ailleurs les mêmes techniques de preuves qui sont utilisées pour démontrer cette dernière. On verra dans la Section 1.3.3 que cette borne inférieure est optimale lorsque δ tend vers 0.

Théorème 1.3.1. *Soit $\mathcal{S} \in \{\mathcal{M}, \mathcal{I}\}$ et $\delta \in (0, 1/2]$. Pour tout algorithme δ-correct sur $\mathcal{S}$ et tout problème de bandit $\boldsymbol{\mu} \in \mathcal{S}$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geqslant T^\star_{\mathcal{S}}(\boldsymbol{\mu})\, \mathrm{kl}(\delta, 1-\delta)\,, \tag{1.21}$$

où le temps caractéristique $T^\star_{\mathcal{S}}(\boldsymbol{\mu})$ est donné par

$$T^\star_{\mathcal{S}}(\boldsymbol{\mu})^{-1} = \sup_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a-\lambda_a)^2}{2}\,. \tag{1.22}$$

En particulier, cela implique

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geqslant T^\star_{\mathcal{S}}(\boldsymbol{\mu})\,.$$

Ce résultat est une généralisation du Théorème 1 de Garivier and Kaufmann [2016]. En effet le problème classique de l'identification du meilleur bras est un cas particulier du cadre non-croissant $\mathcal{S}=\mathcal{M}$ avec un seuil infini $S=+\infty$.

Esquisse de preuve. Cette preuve est quasiment identique à celle du Théorème 1 de Garivier and Kaufmann [2016]. On fixe un problème $\boldsymbol{\mu}$ et une alternative $\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})$. Toujours par contraction de l'entropie (cf. Lemme 1.4.1) on a

$$\mathbb{E}[\tau_\delta]\sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}\big[N_a(\tau_\delta)\big]}{\mathbb{E}[\tau_\delta]} \frac{(\mu_a-\lambda_a)^2}{2} \geqslant \mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}}), \mathbb{P}_{\boldsymbol{\lambda}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}})\big)\,.$$

En remarquant que puisque la stratégie est δ-correct

$$\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}}) \leqslant \delta \leqslant \frac{1}{2} \leqslant 1-\delta \leqslant \mathbb{P}_{\boldsymbol{\lambda}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}})\,,$$

en utilisant des propriétés de monotonie de kl, on obtient

$$\mathrm{kl}\big(\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}}), \mathbb{P}_{\boldsymbol{\lambda}}(\widehat{a}_{\tau_\delta}\neq a^\star_{\boldsymbol{\mu}})\big) \geqslant \mathrm{kl}(\delta,1-\delta)\,.$$

Puis, en passant à l'infimum pour $\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})$, il vient

$$\mathbb{E}[\tau_\delta] \inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}\big[N_a(\tau_\delta)\big]}{\mathbb{E}[\tau_\delta]} \frac{(\mu_a-\lambda_a)^2}{2} \geqslant \mathrm{kl}(\delta,1-\delta)\,.$$

On conclut en notant que $\Big(\mathbb{E}_{\boldsymbol{\mu}}\big[N_a(\tau_\delta)\big]/\mathbb{E}[\tau]\Big)_{a\in\{1,\dots,K\}} \in \Sigma_K$ puis en passant au supremum pour $\omega\in\Sigma_K$. □

À l'instar de Garivier and Kaufmann [2016], on peut montrer que le supremum de (1.22) est atteints en un unique point et on note $\omega^\star(\boldsymbol{\mu})$ ces poids optimaux

$$\omega^\star(\boldsymbol{\mu}) := \operatorname*{arg\,max}_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a-\lambda_a)^2}{2}\,. \tag{1.23}$$

La définition des temps caractéristiques permet de vérifier que l'on a bien $T^\star_{\mathcal{M}}(\boldsymbol{\mu}) \geqslant T^\star_{\mathcal{I}}(\boldsymbol{\mu})$. On dispose même d'une formule explicite des temps caractéristiques lorsque $K=2$, ce n'est plus le cas pour $K\geqslant 3$.

Proposition 1.3.2. *Pour $K = 2$,*

$$T^\star_{\mathcal{I}}(\boldsymbol{\mu})^{-1} = \frac{(2S - \mu_1 - \mu_2)^2}{8}, \tag{1.24}$$

$$T^\star_{\mathcal{M}}(\boldsymbol{\mu})^{-1} = \frac{\min\big((2S - \mu_1 - \mu_2)^2, (\mu_1 - \mu_2)^2\big)}{8}\,. \tag{1.25}$$

À noter que pour les deux cadres, les poids optimaux sont les poids uniformes $\omega^\star = (1/2, 1/2)$. Cependant l'alternative optimale, i.e. l'élément $\boldsymbol{\lambda}$ de $\overline{\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})}$ (la fermeture de $\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})$) où l'infimum est atteint dans (5.3) pour les poids optimaux $\omega^\star$ n'est pas la même. Si l'ensemble des problèmes considérés est $\mathcal{I}$, l'alternative optimal est $\boldsymbol{\lambda} = \big(S - (\mu_2 - \mu_1)/2,\, S + (\mu_2 - \mu_1)/2\big)$. Autrement dit, dans cette alternative les deux bras sont translatés de tel sorte que le milieu des deux moyennes soit égale au seuil S. Au contraire, si l'on abandonne l'hypothèse de croissance des moyennes, i.e., on se place dans $\mathcal{M}$, toujours avec $\boldsymbol{\mu} \in \mathcal{I}$, l'alternative optimal peut être de deux formes différentes. Si le seuil S se situe entre les deux moyennes alors l'alternative optimal est la même que précédemment. Sinon c'est la même que dans le problème de l'identification du meilleur bras (voir Garivier and Kaufmann [2016]) : $\boldsymbol{\lambda} = \big((\mu_1 + \mu_2)/2,\, (\mu_1 + \mu_2)/2\big)$. Ainsi, si $\mu_1 \leqslant S \leqslant \mu_2$, les deux temps caractéristiques coïncident, comme on peut le voir dans la Figure 1.4. Lorsque $K \geqslant 3$ c'est un peu plus compliqué, voir la Figure 1.5 et la Section 5.2.2.

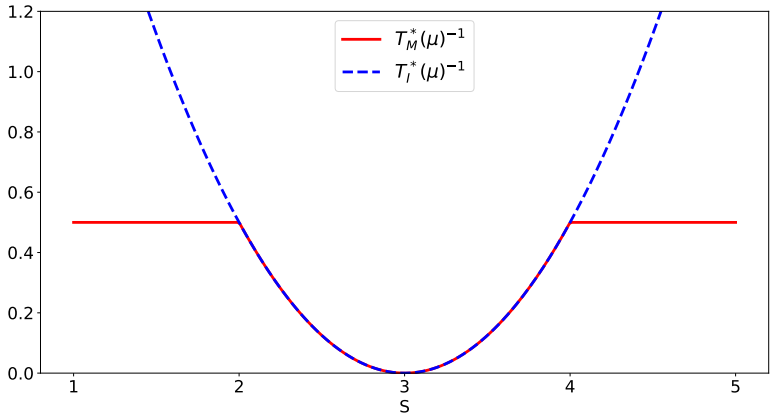

Figure 1.4: L'inverse du temps caractéristique vu comme un fonction du seuil S, pour $\boldsymbol{\mu} = (2, 4)$. En rouge trait plein : cadre non-monotone ($\mathcal{S} = \mathcal{M}$). Bleu pointillé : cadre croissant ($\mathcal{S} = \mathcal{I}$).

1.3.3 Un algorithme asymptotiquement optimal

Un simple adaptation de la procédure *Direct-tracking* de Garivier and Kaufmann [2016] initialement conçue pour le problème de l'identification du meilleur bras donne un algorithme optimal pour le problème de bandit à seuil. Pour tout temps $t \geqslant 1$ soit la fonction

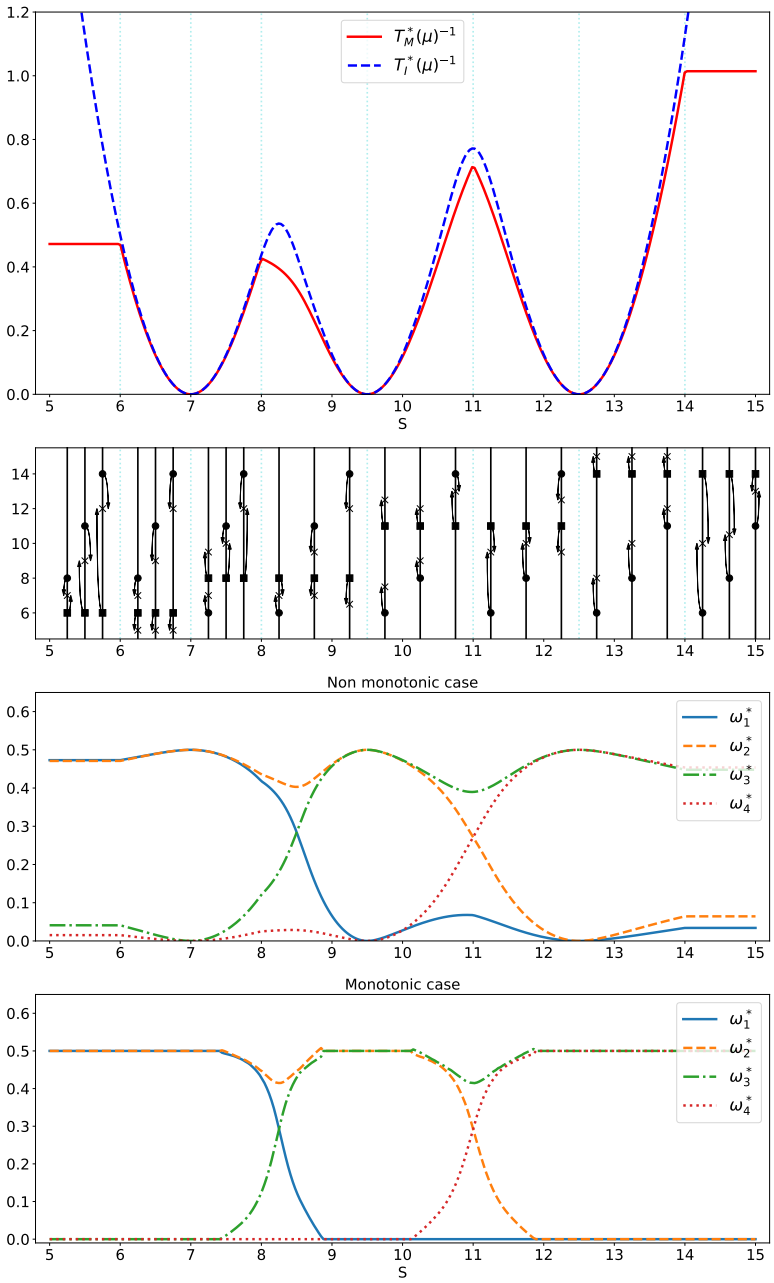

Figure 1.5: *Haut:* L'inverse du temps caractéristique en fonction du seuil S, pour le problème de bandit $\boldsymbol{\mu} = (6, 8, 11, 14)$. En rouge trait plein : cadre non-monotone $\mathcal{S} = \mathcal{M}$. Bleu pointillé : cadre croissant $\mathcal{S} = \mathcal{I}$. *Milieu:* Passage du modèle de bandit initial à l'alternative optimale $\mathcal{M}$. *Bas:* les poids optimaux en fonction du seuil S.

$h(t) = (\sqrt{t} - K/2)_+$ (où $(x)_+$ est la partie positive de x) et soit $U_t = \{a : N_a(t) < h(t)\}$

l'ensemble des bras anormalement peu tirés.

Algorithm 6: Algorithme pour le cas général (Direct-tracking).

Règle d'échantillonnage

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\ N_a(t) & \text{si } U_t \neq \emptyset \quad (\textit{exploration forcée}) \\ \underset{1 \leqslant a \leqslant K}{\operatorname{argmax}}\ t\, w_a^\star(\widehat{\boldsymbol{\mu}}(t)) - N_a(t) & \text{sinon} \quad (\textit{direct tracking}) \end{cases}$$

Critère d'arrêt

$$\tau_\delta = \inf \left\{ t \in \mathbb{N}^* : \widehat{\boldsymbol{\mu}}(t) \in \mathcal{M} \text{ et } \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt\left(\widehat{\boldsymbol{\mu}}(t), \mathcal{S}\right)} \sum_{a=1}^{K} N_a(t) \frac{(\widehat{\mu}_a(t) - \lambda_a)^2}{2} > \beta(t, \delta) \right\}. \tag{1.26}$$

Règle de décision

$$\widehat{a}_{\tau_\delta} \in \underset{1 \leqslant a \leqslant K}{\arg\min} \left| \widehat{\mu}_a(\tau_\delta) - S \right|.$$

Le critère d'arrêt (1.26) ainsi défini permet d'assurer presque indépendemment de la règle d'échantillonnage que l'algorithme est δ-correct. Intuitivement, on s'arrête dès que la divergence de Kullback-Leibler (empirique) entre les moyennes empiriques $\widehat{\boldsymbol{\mu}}(t)$ et celles de l'alternative la plus proche est plus grande qu'un certain seuil $\beta(t, \delta)$. Sinon, à chaque étape l'agent agit comme si les moyennes empiriques $\widehat{\boldsymbol{\mu}}(t)$ étaient égales ou très proches des vraies moyennes $\boldsymbol{\mu}$ en tirant les bras selon les poids optimaux courants $w_a^\star(\widehat{\boldsymbol{\mu}}(t))$. Ce principe est déjà présent et très bien expliqué dans Chernoff [1959]. Cependant on est obligé de forcer l'exploration lorsque certains bras ont été trop peu tirés.

Théorème 1.3.3 (Optimalité asymptotique)**.** *Pour $\mathcal{S} \in \{\mathcal{I}, \mathcal{M}\}$, et une fonction $\beta(t, \delta)$ bien choisie (voir Section 5.3), l'Algorithme 6 est δ-correct sur $\mathcal{S}$ et asymptotiquement optimal, i.e.*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leqslant T_{\mathcal{S}}^\star(\boldsymbol{\mu}). \tag{1.27}$$

La preuve de ce théorème suit celle du Théorème 14 de Garivier and Kaufmann [2016]. Elle est identique pour les deux ensembles d'alternatives $\mathcal{I}$, $\mathcal{M}$ mais l'implémentation pratique varie d'un cadre à l'autre.

Implémentation pratique

La mise en place de l'Algorithme 6 nécessite de calculer efficacement les poids optimaux $w^\star(\boldsymbol{\mu})$ donnés par l'Équation (1.23). Pour le cadre non-monotone $\mathcal{S} = \mathcal{M}$ on peut facilement adapter la procédure de Garivier and Kaufmann [2016, Section 2.2].

Dans le cas croissant $\mathcal{S} = \mathcal{I}$ c'est moins évident. Notons $\mathcal{I}_b := \{\boldsymbol{\lambda} \in \mathcal{I}, a^\star_{\boldsymbol{\lambda}} = b\}$ l'ensemble des alternatives pour lesquelles le bras optimal est b. Puisque la fonction

$$\begin{aligned} F : w \mapsto & \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \\ & = \min_{b \neq a^\star_{\boldsymbol{\mu}}} \inf_{\boldsymbol{\lambda} \in \mathcal{I}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \end{aligned} \tag{1.28}$$

est concave, on peut accéder à son maximum par une montée de gradient sur le simplexe Σ_K. Soit $\overline{\mathcal{I}}_b$ la fermeture de $\mathcal{I}_b$, et soit

$$\boldsymbol{\lambda}^b := \arg\min_{\lambda \in \overline{\mathcal{I}}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \tag{1.29}$$

l'argument de l'infimum (on peut montrer qu'il est unique) dans l'Équation (1.28). Alors, le sous-gradient de F en ω est

$$\partial F(\omega) = \operatorname*{Conv}_{b \in B_{Opt}} \left[\frac{(\mu_a - \lambda^b_a)^2}{2} \right]_{a \in \{1, \cdots, K\}},$$

où Conv désigne l'enveloppe convexe et B_{Opt} l'ensemble des bras qui atteignent le minimum dans la définition (1.28) de F. Ainsi pour pouvoir effectuer la montée de gradient il suffit de résoudre efficacement le problème d'optimisation (1.29). Il se trouve que ce dernier se réduit à une simple régression unimodale (voir Section 5.3.1). Un problème étroitement lié à celui de la régression isotonique, voir par exemple Barlow et al. [1973] et Robertson et al. [1988]). D'ailleurs cette dernière se calcule efficacement via des régressions isotoniques (e.g. Frisén [1986], Geng and Shi [1990], Mureika et al. [1992]) avec une complexité proportionnelle au nombre de bras K. Puisque l'on doit faire cela pour chaque bras $b \neq a^\star_{\boldsymbol{\mu}}$, la complexité pour calculer un sous-gradient est de l'ordre de K^2.

1.3.4 Perspectives

Plusieurs questions restent ouvertes. Une première piste consisterait à généraliser les différents résultats à une famille exponentielle à un paramètre quelconque, par exemple des lois de Bernoulli. Certains semblent se généraliser sans difficultés tels que la borne inférieure (Théorème 1.3.1) ou l'optimalité asymptotique de l'Algorithme 6 (Théorème 1.3.3). Pour d'autres, s'appuyant fortement sur le caractère gaussien, c'est moins évident, comme le calcul effectif des poids optimaux, voir la Section 1.3.3.

Une autre voie serait d'étendre l'étude pour des valeurs modérées de δ (et non se limiter à l'asymptotique $\delta \to 0$) dans la lignée des travaux de Simchowitz et al. [2017]. Un angle d'attaque intéressant serait de mener une étude minimax de ce même problème, cela permettrait entre autres de mieux comprendre la dépendance en K de la complexité qui est un peu cachée dans l'étude asymptotique.

Enfin d'un point de vue plus pratique, il serait intéressant d'utiliser l'algorithme de régression unimodale de Stout [2000] pour calculer directement le sous-gradient de F (Équation (1.28)) avec une complexité en $O(K)$ au lieu de $O(K^2)$.

1.4 Inégalité de Fano

On a vu que l'obtention de bornes inférieures était une étape clé pour l'étude des problèmes de bandits que ce soit pour l'étude du regret avec le Théorème 1.1.2 ou les problèmes de bandit à seuil avec le Théorème 1.3.1. À chaque fois la procédure consistait à minorer une divergence de Kullback-Leibler bien choisie grâce au principe de contraction de l'entropie, cf. Lemme 1.4.1 pour une version générale de ce dernier. En fait ce principe est valable pour n'importe quelle f-divergence, cf. Lemme 6.8.6, mais la divergence de Kullback-Leibler présente l'immense avantage de pouvoir se tensoriser ce qui permet d'effectuer les calculs, à noter que cela mène aussi à des bornes optimales, du moins pour les régimes auxquels on s'est intéressé. C'est pourquoi il est naturel de se tourner vers l'inégalité de Fano qui repose sur ce même principe. C'est d'ailleurs un outil essentiel pour démontrer des bornes inférieures sur l'erreur minimax dans de nombreux autres problèmes de statistique tels que l'estimation non-paramétrique de densités, la régression et la classification (voir, par exemple, Tsybakov, 2009, Massart, 2007).

L'inégalité de Fano est une inégalité d'information qui permet, en particulier, de construire une borne inférieure sur l'erreur minimax dans les problèmes de test d'hypothèses multiples. Elle a aussi d'importantes conséquences en théorie de l'information (voir [Cover and Thomas, 2006]) et dans les domaines adjacents.

Dans un premier temps on présentera une preuve unifiée de plusieurs inégalités de type Fano en mettant en avant la généralité de la démarche adoptée. Puis nous montrerons comment généraliser ces dernières en s'appuyant sur le Lemme 1.1.3 utilisé pour démontrer le Théorème 1.1.2.

1.4.1 Une méthode pour obtenir des inégalités de type Fano

De multiples versions de l'inégalité de Fano coexistent selon le domaine avec lequel on l'aborde. Nous présenterons d'abord une version très générale de cette dernière puis nous énoncerons la version la plus couramment utilisée en statistique. L'obtention de ce type d'inégalité repose sur deux arguments clés : la réduction à des lois de Bernoulli et une borne inférieure sur la divergence kl entre deux lois de Bernoulli. Pour cela, on a besoin de deux outils de la théorie de l'information : la contraction de l'entropie et un de ses corollaire, la convexité jointe de la divergence de Kullback-Leibler.

Lemme 1.4.1 (Contraction de l'entropie)**.** *Soit $\mathbb{P}$ et $\mathbb{Q}$ deux lois de probabilité définies sur le même espace mesurable $(\Omega, \mathcal{F})$, et soit X une variable aléatoire définie sur $(\Omega, \mathcal{F})$. On pose $\mathbb{P}^X$ et $\mathbb{Q}^X$ les mesures images respectivement de $\mathbb{P}$ et $\mathbb{Q}$ par X. Alors,*

$$\mathrm{KL}\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q})\,.$$

Corollaire 1.4.2 (Convexité jointe de KL)**.** *La divergence de Kullback-Leibler* KL *est conjointement convexe, i.e., pour toute mesures de probabilité $\mathbb{P}_1, \mathbb{P}_2$ et $\mathbb{Q}_1, \mathbb{Q}_2$ définies sur le même espace mesurable $(\Omega, \mathcal{F})$, et tout $\lambda \in (0,1)$,*

$$\mathrm{KL}\big(\lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2,\, \lambda\mathbb{Q}_1 + (1-\lambda)\mathbb{Q}_2\big) \leqslant \lambda\,\mathrm{KL}(\mathbb{P}_1, \mathbb{Q}_1) + (1-\lambda)\,\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2)\,.$$

Soit des couples de lois de probabilité $\mathbb{P}_i$, $\mathbb{Q}_i$ et des évènements A_i (pas nécessairement disjoints), où $i \in \{1, \dots, N\}$ avec $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$. L'objectif est de majorer $(1/N)\sum_{i=1}^N \mathbb{P}_i(A_i)$ en fonction des divergences $\mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)$ et des probabilités $\mathbb{Q}(A_i)$. Comme annoncé précédemment la première étape consiste à se ramener à des lois de Bernoulli. En utilisant d'abord la convexité jointe de la divergence de Kullback-Leibler (Corollaire 1.4.2), puis la contraction de l'entropie (Lemme 1.4.1) avec la variable aléatoire $X = \mathbb{1}_{A_i}$, on obtient

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i),\, \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{1.30}$$

Que l'on réécrit $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$ en notant

$$\overline{p} = \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \qquad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) \qquad \overline{K} = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{1.31}$$

La seconde et dernière étape consiste à minorer $\mathrm{kl}\big(\overline{p}, \overline{q}\big)$ afin d'obtenir une borne supérieure sur $\overline{p}$. En notant que $\overline{p}\log\big(\overline{p}\big) + \big(1-\overline{p}\big)\log\big(1-\overline{p}\big) \geqslant -\log(2)$, on a, par définition de $\mathrm{kl}\big(\overline{p}, \overline{q}\big)$,

$$\mathrm{kl}\big(\overline{p}, \overline{q}\big) \geqslant \overline{p}\log\big(1/\overline{q}\big) - \log(2)\,, \qquad \text{ainsi} \qquad \overline{p} \leqslant \frac{\mathrm{kl}\big(\overline{p}, \overline{q}\big) + \log(2)}{\log\big(1/\overline{q}\big)}\,. \tag{1.32}$$

En utilisant la majoration $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$ dans (6.4) on vient de prouver la proposition suivante :

Proposition 1.4.3. *Pour tout couples de lois de probabilité $\mathbb{P}_i$, $\mathbb{Q}_i$ et évènements A_i (non nécessairement disjoints) définis sur le même espace mesurable $(\Omega, \mathcal{F})$, où $i \in \{1, \dots, N\}$, avec $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, on a*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i) + \log(2)}{-\log\left(\dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathbb{Q}_i(A_i)\right)}\,. \tag{1.33}$$

Si $N \geqslant 2$, on peut retrouver la version classique de l'inégalité de Fano en prenant pour les $(A_i)_{1\leqslant i\leqslant N}$ une partition de l'espace sous-jacent et $\mathbb{Q}_i = \mathbb{Q}$ pour tout $1 \leqslant i \leqslant N$:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N}\inf\limits_{\mathbb{Q}}\displaystyle\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \log(2)}{\log(N)}\,,$$

où l'infimum porte sur toutes les lois de probabilité définies sur Ω. Cette inégalité est triviale lorsque $N \leqslant 2$. À noter que l'Inégalité (1.33) permet de mieux comprendre d'où

vient le facteur $\log(N)$ dans l'inégalité de Fano classique. Une autre version, due à Birgé [2005] et Massart [2007] est aussi très populaire parmi les statisticiens : avec les même notations, toujours si les $(A_i)_{1\leqslant i\leqslant N}$ forment une partition et $N \geqslant 2$,

$$\min_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{c,\ \frac{\overline{K}}{\log(N)}\right\} \qquad \text{où} \qquad \overline{K} = \frac{1}{N-1}\sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1) \tag{1.34}$$

pour une certaine constante $c \in (0,1)$. Elle se prouve de la même façon que précédemment en prenant $\mathbb{Q}_i = \mathbb{P}_1$ pour tout i et en modifiant légèrement le passage avec l'Inégalité (1.32), voir le Théorème 6.6.3 et sa preuve.

Dans les deux cas, le lien avec le test d'hypothèses multiples est le suivant : lorsque l'on prend des évènements de la forme $A_i = \{\widehat{\theta} = i\}$, ces deux inégalités fournissent une borne inférieure sur l'erreur minimax $\max_{1\leqslant i\leqslant N} \mathbb{P}_i\big(\widehat{\theta} \neq i\big)$ pour tout estimateur $\widehat{\theta}$, en effet

$$\max_{1\leqslant i\leqslant N} \mathbb{P}_i\big(\widehat{\theta} \neq i\big) = 1 - \min_{1\leqslant i\leqslant N} \mathbb{P}_i\big(\widehat{\theta} = i\big) \geqslant 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i\big(\widehat{\theta} = i\big)\,.$$

1.4.2 Une illustration pour $N = 1$: une preuve du théorème de Cramér pour des lois de Bernoulli

La proposition suivante est un résultat bien connu sur les déviations de la moyenne empirique d'un échantillon de variables indépendantes identiquement distribuées selon une loi de Bernoulli. C'est un cas particulier du théorème de Cramér, voir Cramér [1938], Chernoff [1952], voir aussi Cerf and Petit [2011] pour d'autres références et une preuve dans un contexte bien plus général.

Proposition 1.4.4 (Théorème de Cramér pour des lois de Bernoulli)**.** *Soit $\theta \in (0,1)$. Soit $X_1, \dots, X_n$ i.i.d. selon une loi de Bernoulli* $\mathrm{Ber}(\theta)$*. En posant $\mathbb{P}_\theta$ la mesure de probabilité sous-jacente, on a, pour tout $x \in (\theta, 1)$,*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) = -\mathrm{kl}(x,\theta)\,.$$

La borne supérieure est une simple conséquence de l'inégalité de Chernoff, voir Inégalité (1.4). On pose $\overline{X}_n \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} X_i$. Pour la borne inférieure on va utiliser l'inégalité de Fano. Soit $\varepsilon > 0$ assez petit afin que $x + \varepsilon < 1$. D'après l'Inégalité (1.33) avec la loi $\mathbb{P} = \mathbb{P}_{x+\varepsilon}$ et $\mathbb{Q} = \mathbb{P}_\theta$, et l'évènement $A = \big\{\overline{X}_n > x\big\}$, on a

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) + \log(2)}{\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big)}\right)\,.$$

En notant que $\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) = n\mathrm{kl}(x+\varepsilon, \theta)$ il vient

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant \exp\left(-\frac{n\mathrm{kl}(x+\varepsilon,\theta) + \log 2}{\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big)}\right) \geqslant \exp\left(-\frac{n\mathrm{kl}(x+\varepsilon,\theta) + \log 2}{1 - e^{-n\mathrm{kl}(x,x+\varepsilon)}}\right), \tag{1.35}$$

où la dernière inégalité provient de nouveau de l'inégalité de Chernoff $\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big) = 1 - \mathbb{P}_{x+\varepsilon}\big(\overline{X}_n \leqslant x\big) \geqslant 1 - e^{-n\mathrm{kl}(x,x+\varepsilon)}$. En prenant le logarithme des deux cotés et en passant à la limite $n \to +\infty$ on obtient finalement

$$\liminf_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant -\mathrm{kl}(x+\varepsilon, \theta)\,.$$

On conclut en passant à la limite $\varepsilon \to 0$.

L'inégalité de Fano permet ici d'éviter de faire explicitement un changement de mesure comme c'est le cas dans la preuve historique, ou plutôt l'encapsule dans un outil de plus haut niveau. Cette preuve est assez similaire à celle de la borne inférieure de Lai et Robbins, Théorème 1.1.2, à la différence notable près, que l'on part du problème modifié pour arriver au problème originel, i.e., on a inversé les arguments dans la divergence de Kullback-Leibler.

1.4.3 Extensions de la réduction à des lois de Bernoulli

En reprenant la preuve de la Proposition 1.4.3 on peut ajouter un degré de généralité en utilisant la même astuce qui consiste à se ramener à devoir minorer une divergence de Kullback-Leibler par celle entre deux lois de Bernoulli bien choisies. Plus précisément lorsque l'on parle de réduction à des lois de Bernoulli on fait référence aux Inégalités (1.30). On présente plusieurs extensions possibles.

Distributions indexées par un ensemble éventuellement continu de paramètres. On considère deux modèles statistiques $\mathbb{P}_\theta$, $\mathbb{Q}_\theta$ avec un ensemble mesurable de paramètres $(\Theta, \mathcal{G})$, muni d'une une loi a priori ν sur Θ, et une collection d'évènements A_θ (non nécessairement disjoints) tels que

$$\theta \in \Theta \longmapsto \big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big) \qquad \text{et} \qquad \theta \in \Theta \longmapsto \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

sont $\mathcal{G}$–mesurable. La réductions devient (en utilisant une version généralisée de l'inégalité de Jensen, Lemme 6.8.12) :

$$\begin{aligned} \mathrm{kl}\left(\int_\Theta \mathbb{P}_\theta(A_\theta)\mathrm{d}\nu(\theta),\, \int_\Theta \mathbb{Q}_\theta(A_\theta)\mathrm{d}\nu(\theta)\right) &\leqslant \int_\Theta \mathrm{kl}\big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big)\mathrm{d}\nu(\theta) \\ &\leqslant \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)\mathrm{d}\nu(\theta)\,. \end{aligned} \tag{1.36}$$

Variables aléatoires. Dans les réductions précédentes il n'a jamais été nécéssaire que les évènements A_i ou A_θ forment une partition ou qu'ils soient disjoints. Il n'est donc pas surprenant que l'on puisse remplacer les indicatrices $\mathbb{1}_{A_i}$ ou $\mathbb{1}_{A_\theta}$ utilisées ci-dessus par des variables aléatoires Z_i ou Z_θ à valeurs dans $[0,1]$. La façon la plus élégante de le voir est d'utiliser le Lemme 1.1.3 suivant (utilisé pour prouver la borne de Lai et Robbins, Théorème 1.1.2) qui est une conséquence du Lemme 6.2.2. On énonce la réduction pour

une nombre fini de distributions ainsi que pour un nombre quelconque de distributions indexées par un ensemble éventuellement continu.

Dans le premier cas, on considère une collection $Z_1, \dots, Z_N$ de variables aléatoires à valeurs dans $[0,1]$ et on pose $\mathbb{E}_{\mathbb{P}_i}$ et $\mathbb{E}_{\mathbb{Q}_i}$ leurs espérances selon la loi $\mathbb{P}_i$ et $\mathbb{Q}_i$, la réduction est alors

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{P}_i}\big[Z_i\big],\ \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big]\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\Big(\mathbb{E}_{\mathbb{P}_i}\big[Z_i\big],\ \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big]\Big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{1.37}$$

Dans le cas générale, on pose Z_θ les variables à valeurs dans $[0,1]$, où $\theta \in \Theta$, et où $\mathbb{E}_{\mathbb{P}_\theta}$ et $\mathbb{E}_{\mathbb{Q}_\theta}$ sont l'espérance selon $\mathbb{P}_\theta$ et $\mathbb{Q}_\theta$, on suppose toujours que

$$\theta \in \Theta \longmapsto \Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\ \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\Big) \qquad \text{and} \qquad \theta \in \Theta \longmapsto \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

sont $\mathcal{G}$-mesurable. La réduction est alors

$$\begin{aligned}\mathrm{kl}\left(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta),\ \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta)\right) &\leqslant \int_\Theta \mathrm{kl}\Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\ \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\Big)\mathrm{d}\nu(\theta) \\ &\leqslant \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)\mathrm{d}\nu(\theta)\,. \end{aligned} \tag{1.38}$$

Plusieurs extensions de l'inégalité de Fano ont déjà été développées par le passé. On peut par exemple citer, Han and Verdú [1994] qui ont traité le cas d'un ensemble dénombrable de distributions puis Duchi and Wainwright [2013] et Chen et al. [2016] ont généralisé ce résultat pour des ensembles non-dénombrables de distributions, dans le même esprit que (1.38). Gushchin [2003] ont étendu l'inégalité de Fano dans une autre direction en considérant des variables aléatoires Z_i à valeurs dans $[0,1]$ qui somment à $Z_1 + \ldots + Z_N = 1$, à la place du cas particulier $Z_i = \mathbb{1}_{A_i}$.

1.4.4 Minorations alternatives de kl

Dans la preuve de la Proposition 1.4.3 où dans la section précédente, on a montré qu'après la réduction à des lois de Bernoulli, on obtenait une inégalité de la forme ($\overline{p}$ étant le terme à majorer)

$$\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}\,,$$

où $\overline{K}$ est une moyenne de divergences de Kullback-Leibler, et $\overline{p}$ et $\overline{q}$ sont des moyennes de probabilités où des moyennes de variables aléatoires à valeurs dans $[0,1]$.Il s'agit donc de minorer la fonction kl. On vient d'utiliser la minoration suivante pour la preuve de la Proposition 1.4.3, qui est bien connue, voir par exemple Guntuboyina [2011].

La minoration la plus classique. Pour tout $p \in [0,1]$ et $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\log(1/q) - \log(2)\,, \qquad \text{ainsi} \qquad p \leqslant \frac{\mathrm{kl}(p,q) + \log(2)}{\log(1/q)}\,. \tag{1.39}$$

Une conséquence de l'inégalité de convexité. Cette borne est déjà connue, voir par exemple Chen et al. [2016]. Pour tout $p \in [0,1]$ et $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\log(1/q) - \log(2-q)\,, \qquad \text{ainsi} \qquad p \leqslant \frac{\mathrm{kl}(p,q) + \log(2-q)}{\log(1/q)}\,. \tag{1.40}$$

Enfin on peut citer la borne suivante, issue d'une inégalité de Pinsker raffinée (voir Théorème 6.7.1 due à Weissman et al. [2003]).

Une conséquence d'une inégalité de Pinsker raffinée. Pour tout $p \in [0,1]$ et $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant \max\left\{\log\left(\frac{1}{q}\right), 2\right\}(p-q)^2\,, \qquad \text{ainsi} \qquad p \leqslant q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\{\log(1/q), 2\}}}\,. \tag{1.41}$$

On remarque que cette inégalité interpole entre l'inégalité de Fano classique (Inégalité (1.39)) lorsque l'on minore le maximum par $\log(1/q)$ et l'inégalité de Pinsker lorsque l'on minore le maximum par 2.

1.4.5 Inégalités de Fano généralisées

En combinant les résultats des Sections 1.4.3 et 1.4.4 on obtient des versions généralisées de l'inégalité de Fano. Par, exemple en associant les Inégalités (1.37) et (1.41) on aboutit à une inégalité de type Fano pour un nombre fini de variables aléatoires qui ne somment pas nécessairement à 1.

Lemma 1.4.5. *Pour toutes paires de lois de probabilité $\mathbb{P}_i$, $\mathbb{Q}_i$ et pour toutes variables aléatoires Z_i à valeurs dans $[0,1]$ définies sur le même espace mesurable sous-jacent, où $i \in \{1,\dots,N\}$, avec*

$$0 < \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}[Z_i] < 1\,,$$

on a

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}_i}[Z_i] \leqslant \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}[Z_i] + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)}{-\log\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}[Z_i]\right)}}\,.$$

Cette borne peut par exemple servir à prouver de manière élémentaire la borne inférieure asymptotique de Kwon and Perchet [2016] sur le regret pour des problèmes de bandit adversarial avec des pertes creuses (voir Kwon and Perchet [2016] et Cesa-Bianchi et al. [1997] pour une introduction à ce type de problèmes de bandit). L'énoncer de cette borne ainsi que sa preuve sont donnés dans la Section 6.4.2. On peut aussi combiner les Inégalités (1.38) et (1.39) pour obtenir une version continue de l'inégalité de Fano. (On ne se préoccupe pas ici des questions de mesurabilité.)

Lemma 1.4.6. *On considère un espace mesurable $(\Theta, \mathcal{E})$ muni d'une distribution ν. Étant donné un espace mesurable sous-jacent $(\Omega, \mathcal{F})$, pour toute collection de paires $(\mathbb{P}_\theta, \mathbb{Q}_\theta)$, de lois de probabilité définies sur cet espace sous-jacent et toutes variables aléatoires Z_θ définies sur $(\Omega, \mathcal{F})$, où $\theta \in \Theta$, avec*

$$0 < \int_\Theta \mathbb{E}_{Q_\theta}[Z_\theta] \mathrm{d}\nu(\theta) < 1\,,$$

on a

$$\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big] \mathrm{d}\nu(\theta) \leqslant \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big] \mathrm{d}\nu(\theta) + \sqrt{\frac{\displaystyle\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta) \mathrm{d}\nu(\theta)}{\displaystyle -\log \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big] \mathrm{d}\nu(\theta)}}\,.$$

Cette inégalité permet de montrer, ici aussi de façon très simple, une borne inférieure sur la vitesse de concentration de l'a posteriori dans un contexte bayésien, voir Hoffmann et al. [2015]. Cette borne est prouvée en Section 6.4.1.

1.4.6 Perspectives

Un prolongement naturel de ce travail est de trouver de nouvelles applications aux outils-méthodes présentés, autant pour les problèmes de bandit stochastique que dans d'autres domaines de la statistique. On pourrait, par exemple, penser à des problèmes de bandit plus complexes comme les bornes inférieures pour les problèmes de bandit linéaire. Il semblerait, a priori, que l'apport soit purement esthétique.

Un autre point important serait de comprendre en profondeur, pourquoi dans certaines preuves de bornes inférieures la divergence de Kullback-Leibler apparaît dans un certain sens ou dans l'autre. Par exemples dans la preuve de la borne inférieure de Lai et Robbins (Théorème 1.1.2) on part du problème initial et l'on effectue de manière implicite un changement de mesure vers un problème alternatif. Tandis que pour la preuve de la Proposition 1.4.4 on part de l'alternative pour revenir au problème initial, ces deux changements n'étant pas équivalents.

Chapter 2

Explore First, Exploit Next: The True Shape of Regret in Bandit Problems

In collaboration with Aurélien Garivier and Gilles Stoltz.

Contents

2.1 Introduction.

After the works of Lai and Robbins [1985] and Burnetas and Katehakis [1996], it is widely admitted that the growth of the cumulative regret in a bandit problem is a logarithmic function of time, multiplied by a sum of terms involving Kullback-Leibler divergences. The asymptotic nature of the lower bounds, however, appears clearly in numerical experiments, where the logarithmic shape is not to be observed on small horizons (see Figure 2.1, left). Even on larger horizons, the second-order terms keep a large importance, which causes the regret of some algorithms to remain way *below* the "lower bound" on any experimentally visible horizon (see Figure 2.1, right; see also Garivier et al. [2016]).

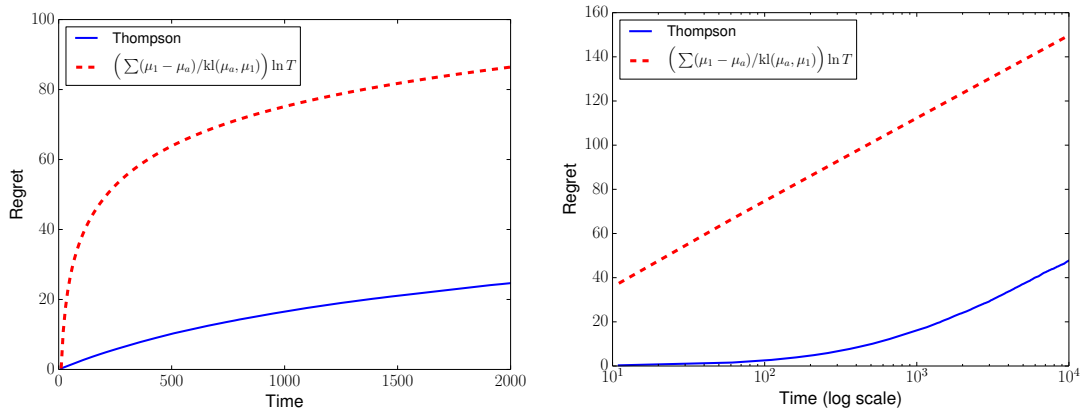

Figure 2.1: Expected regret of Thompson [1933] Sampling (*blue, solid* line) on a Bernoulli bandit problem with parameters $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$; expectations are approximated over 500 runs.

Versus the Lai and Robbins [1985] lower bound (*red, dotted* line) for a Bernoulli model; here kl denotes the Kullback-Leibler divergence (2.5) between Bernoulli distributions.

Left: the shape of regret is not logarithmic at first, rather linear.
Right: the asymptotic lower bound is out of reach unless T is extremely large.

First contribution: a folk result made rigorous. It seems to be a folk result (or at least, a widely believed result) that the regret should be linear in an initial phase of a bandit problem. However, all references that we were pointed out exhibit such a linear behavior only for limited bandit settings; we discuss them below, in the section about literature review. We are the first to provide linear distribution-dependent lower bounds for small horizons that hold for general bandit problems, with no restriction on the shape or on the expectations of the distributions over the arms.

Thus we may draw a more precise picture of the behavior of the regret in any bandit problem. Indeed, our bounds show the existence of three successive phases: an initial

linear phase, when all the arms are essentially drawn uniformly; a transition phase, when the number of observations becomes sufficient to perceive differences; and the final phase, when the distributions associated with all the arms are known with high confidence and when the new draws are just confirming the identity of the best arms with higher and higher degree of confidence (this is the famous logarithmic phase). This last phase may often be out of reach in applications, especially when the number of arms is large.

Second contribution: a generic tool for proving distribution-dependent bandit lower bounds. On the technical side, we provide simple proofs, based on the fundamental information-theoretic inequality (2.6) stated in Section 2.2, which generalizes and simplifies previous approaches based on explicit changes of measures. In particular, we are able to re-derive the asymptotic distribution-dependent lower bounds of Lai and Robbins [1985], Burnetas and Katehakis [1996] and Cowan and Katehakis [2015] in a few lines. This may perhaps be one of the most striking contributions of this chapter. As a final set of results, we offer non-asymptotic versions of these lower bounds for large horizons, and exhibit the optimal order of magnitude of the second-order term in the regret bound, namely, $-\log(\log T)$.

The proof techniques come to the essence of the arguments used so far in the literature and they involve no unnecessary complications; they only rely on well-known properties of Kullback-Leibler divergences.

2.1.1 Setting.

We consider the simplest case of a stochastic bandit problem, with finitely many arms indexed by $a \in \{1, \dots, K\}$. Each of these arms is associated with an unknown probability distribution ν_a over $\mathbb{R}$. We assume that each ν_a has a well-defined expectation and call $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ a bandit problem.

At each round $t \geqslant 1$, the player pulls the arm A_t and gets a real-valued reward Y_t drawn independently at random according to the distribution ν_{A_t}. This reward is the only piece of information available to the player.

Strategies. A strategy ψ associates an arm with the information gained in the past, possibly based on some auxiliary randomization; without loss of generality, this auxiliary randomization is provided by a sequence $U_0, U_1, U_2, \dots$ of independent and identically distributed random variables, with common distribution the uniform distribution over $[0, 1]$. Formally, a strategy is a sequence $\psi = (\psi_t)_{t \geqslant 0}$ of measurable functions, each of which associates with the said past information, namely,

$$I_t = \big(U_0,\, Y_1, U_1,\, \dots,\, Y_t, U_t\big)\,,$$

an arm $\psi_t(I_t) = A_{t+1} \in \{1, \dots, K\}$, where $t \geqslant 0$. The initial information reduces to $I_0 = U_0$ and the first arm is $A_1 = \psi_0(U_0)$. The auxiliary randomization is conditionally independent of the sequence of rewards in the following sense: for $t \geqslant 1$, the randomization U_t used to pick A_{t+1} is independent of I_{t-1} and Y_t.

Regret. A typical measure of the performance of a strategy is given by its regret. To recall its definition, we denote by $E(\nu_a) = \mu_a$ the expected payoff of arm a and by Δ_a its gap to an optimal arm:

$$\mu^\star = \max_{a=1,\dots,K} \mu_a \qquad \text{and} \qquad \Delta_a = \mu^\star - \mu_a \,.$$

The number of times an arm a is pulled until round T by a strategy ψ is referred to as

$$N_{\psi,a}(T) = \sum_{t=1}^{T} \mathbb{1}_{\{A_t=a\}} = \sum_{t=1}^{T} \mathbb{1}_{\{\psi_{t-1}(I_{t-1})=a\}} \,.$$

The expected regret of a strategy ψ equals, by the tower rule (see details below),

$$R_{\psi,\underline{\nu},T} = T\mu^\star - \mathbb{E}_{\underline{\nu}}\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}_{\underline{\nu}}\left[\sum_{t=1}^{T} \big(\mu^\star - \mu_{A_t}\big)\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \,. \tag{2.1}$$

In the equation above, the notation $\mathbb{E}_{\underline{\nu}}$ refers to the expectation associated with the bandit problem $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$; it is made formal in Section 2.2.

To show (2.1), we use that by the definition of the bandit setting, the distribution of the obtained payoff Y_t only depends on the chosen arm A_t and is independent from the past random draws of the $Y_1, \dots, Y_{t-1}$. More precisely, conditionally on A_t, the distribution of Y_t is ν_{A_t} so that

$$\mathbb{E}_{\underline{\nu}}\big[Y_t \,|\, A_t\big] = \mu_{A_t} \,, \qquad \text{thus} \qquad \mathbb{E}_{\underline{\nu}}[Y_t] = \mathbb{E}_{\underline{\nu}}\Big[\mathbb{E}_{\underline{\nu}}\big[Y_t \,|\, A_t\big]\Big] = \mathbb{E}_{\underline{\nu}}\big[\mu_{A_t}\big] \,,$$

where we used the tower rule for the second set of equalities.

2.1.2 The general asymptotic lower bound: a quick literature review.

We consider a bandit model $\mathcal{D}$, i.e., a collection of possible distributions ν_a associated with the arms. (That is, $\mathcal{D}$ is a subset of the set of all possible distributions over $\mathbb{R}$ with an expectation.) Lai and Robbins [1985] and later Burnetas and Katehakis [1996] exhibited asymptotic lower bounds and matching asymptotic upper bounds on the normalized regret $R_{\psi,\underline{\nu},T}/\log T$, respectively in a one-parameter case and in a more general, multi-dimensional parameter case, under mild conditions on $\mathcal{D}$. We believe that the extension of these bounds to any, even non-parametric, model was a known or at least conjectured result (see, for instance, the introduction of [Cappé et al., 2013]). It turns out that recently, Cowan and Katehakis [2015] provided a clear non-parametric statement, though under additional mild conditions on the model $\mathcal{D}$, which, as we will see, are not needed.

We recall that we denoted by E the expectation operator (that associates with each distribution its expectation).

To state the bound for the case of an arbitrary model $\mathcal{D}$, we will use the following key quantity $\mathcal{K}_{\inf}$ introduced by Burnetas and Katehakis [1996, quantity (3)–(b) on page 125].

The key quantity $\mathcal{K}_{\inf}$. For any given $\nu_a \in \mathcal{D}$ and any real number x,

$$\mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) : \quad \nu'_a \in \mathcal{D} \quad \text{and} \quad E(\nu'_a) > x\Big\};$$

by convention, the infimum of the empty set equals $+\infty$. When the considered strategy is uniformly fast convergent in the sense of Definition 2.2.4 (stated later in this chapter), then, for any suboptimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\,. \tag{2.2}$$

Note that by the convention on the infimum of the empty set, this lower bound is void as soon as there exists no $\nu'_a \in \mathcal{D}$ such that $E(\nu'_a) > \mu^\star$.

Previous partial simplifications of the proof of (2.2). We re-derive the above bound in a few lines in Section 2.5.2.

There had been recent attempts to clarify the exposition of the proof of this lower bound, together with the desire of dropping the mild conditions that were still needed so far on the model $\mathcal{D}$. We first mention that Cowan and Katehakis [2015] provided a more general and streamlined approach than the original expositions by Lai and Robbins [1985] and Burnetas and Katehakis [1996].

The case of Bernoulli models was discussed in Bubeck [2010] and Bubeck and Cesa-Bianchi [2012]. Only assumptions of uniform fast convergence of the strategies are required (see Definition 2.2.4) and the associated proof follows the original proof technique, by performing first an explicit change of measure and then applying some Markov–Chernoff bounding. More recently, Jiang [2015, Section 2.2] presented a proof (only in the Bernoulli case) not relying on any explicit change of measure but with many additional technicalities with respect to our exposition, including some Markov bounding of well-chosen events. We have been referred to this PhD dissertation only recently.

As far as general bandit models are concerned, we may cite Kaufmann et al. [2016, Appendix B]: they deal with the case of any model $\mathcal{D}$ but with the restriction that only bandit problems $\underline{\nu}$ with a unique optimal arm should be considered. They still use both an explicit change of measure –to prove the chain-rule equality in (2.6)– and then apply as well some Markov–Chernoff bounding to the probability of well-chosen events. With a different aim, Combes and Proutière [2014] presented similar arguments.

We also wish to mention the contribution of Wu et al. [2015], though their focus and aim are radically different. With respect to some aspects, their setting and goal is wider or more general: they developed non-asymptotic problem-dependent lower bounds on the regret of any algorithm, in the case of more general limited feedback models than just the simplest case of multi-armed bandit problems. Their lower bounds can recover the asymptotic bounds of Burnetas and Katehakis [1996], but only up to a constant factor as they acknowledge in their contribution. These lower bounds are in terms of uniform upper bounds on the regret of the considered strategies, which is in contrast with the lower bounds we develop in Section 2.3. Therein, we need some assumptions on

the strategies –extremely mild ones, though: some minimal symmetry– and do not need their regret to be bounded from above. However, the main difference with respect to this reference is that its focus is limited to specific bandit models, namely Gaussian bandits models, while Burnetas and Katehakis [1996] and we do not impose such a restriction on the bandit model.

2.1.3 Other bandit lower bounds: a brief literature review.

Here we are mostly interested in general distribution-dependent lower bounds, that hold for all bandit problems, just like (2.2). We do target generality. This is in contrast with many earlier lower bounds in the multi-armed bandit setting, which are rather of the following form, which we will refer to as (well-chosen):

> "*There exists some* well-chosen*, difficult bandit problem such that all strategies suffer a regret larger than* [...]." (well-chosen)

Specific examples and pointers for this kind of bounds are given below. An interesting variation is provided by Mannor and Tsitsiklis [2004a, Theorem 10], who state that for all strategies, there exists some well-chosen, difficult Bernoulli bandit problem such that the regret is linear at first and then, logarithmic.

On the contrary, we will issue statements of the following form, which we will refer to as (all):

> "*For* all *bandit problems, all (reasonable) strategies suffer a regret larger than* (...)." (all)

Sometimes, but not always, we will have to impose some mild restrictions on the considered strategies (like some minimal symmetry, or some notion of uniform fast convergence); this is what we mean by requiring the strategies to be "reasonable".

We discuss briefly below two other sets of regret lower bounds. We are pleased to mention that our fundamental inequality was already used in at least one subsequent article, namely by Garivier et al. [2016], to prove in a few lines matching lower bounds for a refined analysis of explore-then-commit strategies.

The distribution-free lower bound. This inequality states that for the model $\mathcal{D} = \mathcal{M}\big([0,1]\big)$ of all probability distributions over $[0,1]$, for all strategies ψ, for all $T \geqslant 1$ and all $K \geqslant 2$,

$$\sup_{\underline{\nu}} R_{\psi,\underline{\nu},T} \geqslant \frac{1}{20} \min\Big\{\sqrt{KT},\, T\Big\}\,; \tag{2.3}$$

see Auer et al. [2002b], Cesa-Bianchi and Lugosi [2006], and for two-armed bandits, Kulkarni and Lugosi [2000]. We re-derive the above bound in Section 2.5.2 of the appendix. This re-derivation follows the very same proof scheme as in the original proof; the only difference is that some steps (e.g., the use of chain-rule equality for Kullback-Leibler divergences) are implemented separately as parts of the proof of our general inequality (2.6). In particular, the well-chosen difficult bandit problems used to prove

this bound are composed of Bernoulli distributions with parameters $1/2$ and $1/2+\varepsilon$, where ε is carefully tuned according to the values of T and K. This bound therefore rather falls under the umbrella (well-chosen).

Lower bounds for sub-Gaussian bandit problems in the case when $\mu^\star$ or the gaps Δ are known. This framework and the exploitation of this knowledge was first studied by Bubeck et al. [2013a]. They consider a bandit model $\mathcal{D}$ containing only sub-Gaussian distributions with parameter $\sigma^2 \leqslant 1$; that is, distributions ν_a, with expectations $\mu_a \in \mathbb{R}$, such that

$$\forall \lambda \in \mathbb{R}, \qquad \int_{\mathbb{R}} \exp\bigl(\lambda(y-\mu_a)\bigr)\,\mathrm{d}\nu_a(y) \leqslant \exp\left(\frac{\lambda^2}{2}\right). \tag{2.4}$$

Examples of such distributions include Gaussian distributions with variance smaller than 1 and bounded distributions with range smaller than 2.

They study how much smaller the regret bounds can get when either the maximal expected payoff $\mu^\star$ or the gaps Δ_a are known. For the case when the gaps Δ_a are known but not $\mu^\star$, they exhibit a lower bound on the regret matching previously known upper bounds, thus proving their optimality. For the case when $\mu^\star$ is known but not the gaps, they offer an algorithm and its associated regret upper bound, as well as a framework for deriving a lower bound; later work (see [Bubeck et al., 2013b] and [Faure et al., 2015]) point out that a bounded regret can be achieved in this case.

We (re-)derive these two lower bounds in a few lines in Section 2.5.3 of the appendix. In particular, the well-chosen difficult bandit problems used are composed of Gaussian distributions $\mathcal{N}(\mu_a, 1)$, with expectations $\mu_a \in \{-\Delta, 0, \Delta\}$. Only statements of the form (well-chosen), not of the form (all), are obtained. Put differently, no general distribution-dependent statement like: "For all bandit problems in which the gaps Δ (or the maximal expected payoff $\mu^\star$) are known, all (reasonable) strategies suffer a regret larger than [...]" is proposed by Bubeck et al. [2013a]; only well-chosen, difficult bandit problems are considered. This is in strong contrast with our general distribution-dependent bounds for the initial linear regime, provided in Section 2.3.

2.1.4 Outline of our contributions.

In Section 2.2, we present Inequality (2.6), in our opinion the most efficient and most versatile tool for proving lower bounds in bandit models. We carefully detail its remarkably simple proof, together with an elegant re-derivation of the earlier asymptotic lower bounds by Lai and Robbins [1985], Burnetas and Katehakis [1996] and Cowan and Katehakis [2015]. Some other earlier bounds are also re-derived in Appendix 2.5.2, namely, the distribution-free lower bound by Auer et al. [2002b] as well as the bounded-regret Gaussian lower bounds by Bubeck et al. [2013a] in the case when $\mu^\star$ or the gaps Δ are known.

The true power of Inequality (2.6) is illustrated in Section 2.3: we study the initial regime when the small number T of draws does not yet permit to unambiguously identify

the best arm. We propose three different bounds (each with specific merits). They explain the quasi-linear growth of the regret in this initial phase. We also discuss how the length of the initial phase depends on the number of arms and on the gap between optimal and sub-optimal arms in Kullback-Leibler divergence. These lower bounds are extremely strong as they hold for all possible bandit problems, not just for some well-chosen ones.

Section 2.4 contains a general non-asymptotic lower bound for the logarithmic (large T) regime. This bound does not only contain the right leading term, but the analysis aims at highlighting what the second-order terms depend on. Results of independent interest on the regularity (upper semi-continuity) of $\mathcal{K}_{\inf}$ are provided in its Subsection 2.4.2.

2.2 The fundamental inequality, and re-derivation of earlier lower bounds.

We recall that kl denote the Kullback-Leibler divergence for Bernoulli distributions:

$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \,. \tag{2.5}$$

We show in this section that for all strategies ψ, for all bandit problems $\underline{\nu}$ and $\underline{\nu}'$, for all $\sigma(I_T)$–measurable random variables Z with values in $[0,1]$,

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \, \mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z], \, \mathbb{E}_{\underline{\nu}'}[Z]\big) \,. \tag{2.6}$$

Inequality (2.6) will be referred to as the fundamental inequality of this chapter. We will typically apply it by considering variables of the form $Z = N_{\psi,k}(T)/T$ for some arm k. That the kl term in (2.6) then also contains expected numbers of draws of arms will be very handy. Unlike all previous proofs of distribution-dependent lower bounds for bandit problems, we will not have to introduce well-chosen events and control their probability by some Markov–Chernoff bounding. Implicit changes of measures will however be performed by considering bandit problems $\underline{\nu}$ and $\underline{\nu}'$ and their associated probability measures $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$.

Underlying probability measures. The proof of (2.6) will be based, among others, on an application of the chain rule for Kullback-Leibler divergences. For this reason, it is helpful to construct and define the underlying measures, so that the needed stochastic transition kernels appear clearly.

By Kolmogorov's extension theorem, there exists a measurable space $(\Omega, \mathcal{F})$ on which all probability measures $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$ considered above can be defined; e.g., $\Omega = [0,1] \times \big(\mathbb{R} \times [0,1]\big)^{\mathbb{N}}$. Given the probabilistic and strategic setting described in Section 2.1.1, the probability measure $\mathbb{P}_{\underline{\nu}}$ over this $(\Omega, \mathcal{F})$ is such that for all $t \geqslant 0$, for all Borel sets $B \subseteq \mathbb{R}$ and $B' \subseteq [0,1]$,

$$\mathbb{P}_{\underline{\nu}}\big(Y_{t+1} \in B, \, U_{t+1} \in B' \;\big|\; I_t\big) = \nu_{\psi_t(I_t)}(B) \; \lambda(B') \,, \tag{2.7}$$

where λ denotes the Lebesgue measure on $[0,1]$.

Remark 2.2.1. Equation (2.7) actually reveals that the distributions $\mathbb{P}_{\underline{\nu}}$ should be indexed as well by the considered strategy ψ. Because the important element in the proofs will be the dependency on $\underline{\nu}$ (we will replace $\underline{\nu}$ by alternative bandit problems $\underline{\nu}'$), we drop the dependency on ψ in the notation for the underlying probability measures. This will not come at the cost of clarity as virtually all events A_ψ and random variables Z_ψ that will be considered will depend on ψ: we will almost always deal with probabilities of the form $\mathbb{P}_{\underline{\nu}}(A_\psi)$ or expectations of the form $\mathbb{E}_{\underline{\nu}}[Z_\psi]$.

2.2.1 Proof of the fundamental inequality (2.6).

We let $\mathbb{P}_{\underline{\nu}}^{I_T}$ and $\mathbb{P}_{\underline{\nu}'}^{I_T}$ denote the respective distributions (pushforward measures) of I_T under $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$. We add an intermediate equation in (2.6),

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_T},\,\mathbb{P}_{\underline{\nu}'}^{I_T}\big) \geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z],\,\mathbb{E}_{\underline{\nu}'}[Z]\big)\,, \tag{2.8}$$

and are left with proving a standard equality (via the chain rule for Kullback-Leibler divergences) and a less standard inequality (following from the data-processing inequality for Kullback-Leibler divergences).

Remark 2.2.2. Although this possibility is not used in the present chapter, it is important to note, after Kaufmann et al. [2016, Lemma 1], that (2.8) actually holds not only for deterministic values of T but also for any stopping time with respect to the filtration generated by $(I_t)_{t\geqslant 1}$.

Proof of the equality in (2.8). This equality can be found, e.g., in the proofs of the distribution-free lower bounds on the bandit regret, in the special case of Bernoulli distributions, see Auer et al. [2002b] and Cesa-Bianchi and Lugosi [2006]; see also Combes and Proutière [2014]. We thus reprove this equality for the sake of completeness only.

We use the symbol $\otimes$ to denote products of measures. The stochastic transition kernel (2.7) exactly indicates that the conditional distribution of (Y_{t+1},U_{t+1}) given I_t equals

$$\mathbb{P}_{\underline{\nu}}^{(Y_{t+1},U_{t+1})\,|\,I_t} = \nu_{\psi_t(I_t)} \otimes \lambda\,.$$

Because the conditional distribution at hand takes such a simple form, the chain rule for Kullback-Leibler divergences applies; it ensures that for all $t \geqslant 0$,

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_{t+1}},\,\mathbb{P}_{\underline{\nu}'}^{I_{t+1}}\Big) &= \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{(I_t,Y_{t+1},U_{t+1})},\,\mathbb{P}_{\underline{\nu}'}^{(I_t,Y_{t+1},U_{t+1})}\Big)\\ &= \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_t},\,\mathbb{P}_{\underline{\nu}'}^{I_t}\big) + \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{(Y_{t+1},U_{t+1})\,|\,I_t},\,\mathbb{P}_{\underline{\nu}'}^{(Y_{t+1},U_{t+1})\,|\,I_t}\Big)\,,\end{aligned} \tag{2.9}$$

where

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{(Y_{t+1},U_{t+1})\,|\,I_t},\,\mathbb{P}_{\underline{\nu}'}^{(Y_{t+1},U_{t+1})\,|\,I_t}\Big) &= \mathbb{E}_{\underline{\nu}}\Big[\mathbb{E}_{\underline{\nu}}\Big[\mathrm{KL}\big(\nu_{\psi_t(I_t)}\otimes\lambda,\,\nu'_{\psi_t(I_t)}\otimes\lambda\big)\,\Big|\,I_t\Big]\Big]\\ &= \mathbb{E}_{\underline{\nu}}\Big[\mathbb{E}_{\underline{\nu}}\Big[\mathrm{KL}\big(\nu_{\psi_t(I_t)},\,\nu'_{\psi_t(I_t)}\big)\,\Big|\,I_t\Big]\Big]\\ &= \mathbb{E}_{\underline{\nu}}\Big[\sum_{a=1}^{K}\mathrm{KL}(\nu_a,\nu'_a)\,\mathbb{1}_{\{\psi_t(I_t)=a\}}\Big]\,.\end{aligned}$$

Recalling that $A_{t+1}=\psi_t(I_t)$, we proved so far

$$\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_{t+1}},\,\mathbb{P}_{\underline{\nu}'}^{I_{t+1}}\Big)=\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_t},\,\mathbb{P}_{\underline{\nu}'}^{I_t}\big)+\mathbb{E}_{\underline{\nu}}\Big[\sum_{a=1}^{K}\mathrm{KL}(\nu_a,\nu'_a)\,\mathbb{1}_{\{A_{t+1}=a\}}\Big]\,.$$

Iterating the argument and using that $\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_0},\,\mathbb{P}_{\underline{\nu}'}^{I_0}\big)=\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{U_0},\,\mathbb{P}_{\underline{\nu}'}^{U_0}\big)=\mathrm{KL}(\lambda,\lambda)=0$ leads to the equality stated in (2.8).

Proof of the inequality in (2.8). *This is our key contribution to a simplified proof of the lower bound (2.2).* It is a consequence of the data-processing inequality (also known as contraction of entropy), i.e., the fact that Kullback-Leibler divergences between pushforward measures are smaller than the Kullback-Leibler divergences between the original probability measures; see Lemma 2.5.1 in Appendix 2.5.1 for a statement and elements of proof.

We actually state our inequality in a slightly more general way, as it is of independent interest.

Lemma 2.2.3. *Consider a measurable space $(\Gamma,\mathcal{G})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, and any $\mathcal{G}$–measurable random variable $Z:\Omega\to[0,1]$. We denote respectively by $\mathbb{E}_1$ and $\mathbb{E}_2$ the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$. Then,*

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2)\geqslant\mathrm{kl}\big(\mathbb{E}_1[Z],\mathbb{E}_2[Z]\big)\,.$$

Proof. We augment the underlying measurable space into $\Gamma\times[0,1]$, where $[0,1]$ is equipped with the Borel σ–algebra $\mathrm{Ber}\big([0,1]\big)$ and the Lebesgue measure λ. We denote by $\mathcal{G}\otimes\mathrm{Ber}\big([0,1]\big)$ the σ–algebra generated by product sets in $\mathcal{G}\times\mathrm{Ber}\big([0,1]\big)$. Now, for any event $E\in\mathcal{G}\otimes\mathrm{Ber}\big([0,1]\big)$, by the consideration of product distributions for the equality and by the data-processing inequality (Lemma 2.5.1) applied to $X=\mathbb{1}_E$ for the inequality, we have

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2)=\mathrm{KL}\big(\mathbb{P}_1\otimes\lambda,\,\mathbb{P}_2\otimes\lambda\big)\geqslant\mathrm{KL}\Big((\mathbb{P}_1\otimes\lambda)^{\mathbb{1}_E},\,(\mathbb{P}_2\otimes\lambda)^{\mathbb{1}_E}\Big)\,.$$

The distribution $(\mathbb{P}_j\otimes\lambda)^{\mathbb{1}_E}$ of $\mathbb{1}_E$ under $\mathbb{P}_j\otimes\lambda$ is a Bernoulli distribution, with parameter the probability of E under $\mathbb{P}_j\otimes\lambda$; therefore, using the notation kl, we have got so far

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2)\geqslant\mathrm{KL}\Big((\mathbb{P}_1\otimes\lambda)^{\mathbb{1}_E},\,(\mathbb{P}_2\otimes\lambda)^{\mathbb{1}_E}\Big)=\mathrm{kl}\big((\mathbb{P}_1\otimes\lambda)(E),\,(\mathbb{P}_2\otimes\lambda)(E)\big)\,.$$

We consider $E = \big\{(\gamma, x) \in \Gamma \times [0,1] : x \leqslant Z(\gamma)\big\}$ and note noting that for all j, by the Fubini-Tonelli theorem,

$$(\mathbb{P}_j \otimes \lambda)(E) = \int_\Omega \left(\int_{[0,1]} \mathbb{1}_{\{x \leqslant Z(\gamma)\}} \, \mathrm{d}\lambda(x) \right) \mathrm{d}\mathbb{P}_j(\gamma) = \int_\Omega Z(\gamma) \, \mathrm{d}\mathbb{P}_j(\gamma) = \mathbb{E}_j[Z] \, .$$

This concludes the proof of this lemma. □

2.2.2 Application: re-derivation of the general asymptotic distribution-dependent bound.

As a warm-up, we show how the asymptotic distribution-dependent lower bound (2.2) of Burnetas and Katehakis [1996] can be reobtained, for so-called uniformly fast convergent strategies.

Definition 2.2.4. A strategy ψ is uniformly fast convergent on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a, i.e., for all arms a such that $\Delta_a > 0$, for all $0 < \alpha \leqslant 1$, it satisfies $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] = o(T^\alpha)$.

Theorem 2.2.5. *For all models $\mathcal{D}$, for all uniformly fast convergent strategies ψ on $\mathcal{D}$, for all bandit problems $\underline{\nu}$, for all suboptimal arms a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})} \, .$$

Proof. Given any bandit problem $\underline{\nu}$ and any suboptimal arm a, we consider a modified problem $\underline{\nu}'$ where a is the (unique) optimal arm: $\nu'_k = \nu_k$ for all $k \neq a$ and ν'_a is any distribution in $\mathcal{D}$ such that its expectation μ'_a satisfies $\mu'_a > \mu^\star$ (if such a distribution exists; see the end of the proof otherwise). We apply the fundamental inequality (2.6) with $Z = N_{\psi,a}(T)/T$. All Kullback-Leibler divergences in its left-hand side are null except the one for arm a, so that we get the lower bound

$$\begin{aligned}
\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \, \mathrm{KL}(\nu_a, \nu'_a) &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T}, \, \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T} \right) \\
&\geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T}\right) \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]} - \log 2 \, , \qquad (2.10)
\end{aligned}$$

where we used for the second inequality that for all $(p,q) \in [0,1]^2$,

$$\mathrm{kl}(p,q) = \underbrace{p \log \frac{1}{q}}_{\geqslant 0} + (1-p) \log \frac{1}{1-q} + \underbrace{\big(p \log p + (1-p) \log(1-p)\big)}_{\geqslant -\log 2} \, . \qquad (2.11)$$

The uniform fast convergence of ψ together with the fact that all arms $k \neq a$ are suboptimal for $\underline{\nu}'$ entails that

$$\forall \, 0 < \alpha \leqslant 1, \qquad 0 \leqslant T - \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big] = \sum_{k \neq a} \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big] = o(T^\alpha) \, ;$$

in particular, $T - \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big] \leqslant T^\alpha$ for T sufficiently large. Therefore, for all $0 < \alpha \leqslant 1$,

$$\liminf_{T\to\infty} \frac{1}{\log T} \log \frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]} \geqslant \liminf_{T\to\infty} \frac{1}{\log T} \log \frac{T}{T^\alpha} = (1-\alpha)\,.$$

In addition, the uniform fast convergence of ψ and the suboptimality of a for the bandit problem $\underline{\nu}$ ensure that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]/T \to 0$. Substituting these two facts in (2.10) we proved

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{\log T} \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)}\,.$$

By taking the supremum in the right-hand side over all distributions $\nu'_a \in \mathcal{D}$ with $\mu'_a > \mu^\star$, if at least one such distribution exists, we get the bound of the theorem. Otherwise, $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}) = +\infty$ by a standard convention on the infimum of an empty set and the bound holds as well. □

2.3 Non-asymptotic bounds for small values of T.

We prove three such bounds with different merits and drawbacks. Basically, we expect suboptimal arms to be pulled each about T/K of the time when T is small; when T becomes larger, sufficient information was gained for identifying the best arm, and the logarithmic regime can take place.

The first bound shows that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ is of order T/K as long as T is at most of order $1/\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})$; we call it an absolute lower bound for a suboptimal arm a. Its drawback is that the times T for which it is valid are independent of the number of arms K, while (at least in some cases) one may expect the initial phase to last until $T \approx K/\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})$.

The second lower bound thus addresses the dependency of the initial phase in K by considering a relative lower bound between a suboptimal arm a and an optimal arm $a^\star$. We prove that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)/N_{\psi,a^\star}(T)\big]$ is not much smaller than 1 whenever T is at most of order $K/\mathrm{KL}(\nu_a, \nu_{a^\star})$. Here, the number of arms K plays the expected effect on the length of the initial exploration phase, which should be proportional to K.

The third lower bound is a collective lower bound on all suboptimal arms, i.e., a lower bound on $\sum_{a\notin\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ where $\mathcal{A}^\star(\underline{\nu})$ denotes the set of the $A^\star_{\underline{\nu}}$ optimal arms of $\underline{\nu}$. It is of the desired order $T(1 - A^\star_{\underline{\nu}}/K)$ for times T of the desired order $K/\mathcal{K}^{\max}_{\underline{\nu}}$, where $\mathcal{K}^{\max}_{\underline{\nu}}$ is some Kullback-Leibler divergence.

Minimal restrictions on the considered strategies. We prove these lower bounds under minimal assumptions on the considered strategies: either some mild symmetry (much milder than asking for symmetry under permutation of the arms, see Definition 2.3.3); or the fact that for suboptimal arms a, the number of pulls $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ should decrease as μ_a decreases, all other distributions of arms being fixed (see Definitions 2.3.1 and 2.3.5). These assumptions are satisfied by all well-performing strategies

we could think of: the UCB strategy of Auer et al. [2002a], the KL-UCB strategy of Cappé et al. [2013], Thompson [1933] Sampling, EXP3 of Auer et al. [2002b], etc.

These mild restrictions on the considered strategies are necessary to rule out the irrelevant strategies (e.g., always pull arm 1) that would perform extremely well for some particular bandit problems $\underline{\nu}$. This is because we aim at proving distribution-dependent lower bounds that are valid for all bandit problems $\underline{\nu}$: we prefer to impose the (mild) constraints on the strategies.

Note that the assumption of uniform fast convergence (Definition 2.2.4), though classical and well accepted, is quite strong. Note that it is necessary for a strategy to satisfy some symmetry and to be smarter than the uniform strategy in the limit (not for all T, see Definition 2.3.1) to be uniformly fast convergent. Hence, the class of strategies we consider is essentially much larger than the subset of uniformly fast convergent strategies.

2.3.1 Absolute lower bound for a suboptimal arm.

The uniform strategy is the one that pulls an arm uniformly at random at each round.

Definition 2.3.1. A strategy ψ is smarter than the uniform strategy on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all optimal arms $a^\star$, for all $T \geqslant 1$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a^\star}(T)\big] \geqslant \frac{T}{K}\,.$$

Theorem 2.3.2. *For all models $\mathcal{D}$, for all strategies ψ that are smarter than the uniform strategy on $\mathcal{D}$, for all bandit problems $\underline{\nu}$, for all arms a, for all $T \geqslant 1$,*

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{K}\Big(1 - \sqrt{2T\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\Big)\,.$$

In particular,

$$\forall\, T \leqslant \frac{1}{8\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}, \qquad \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{2K}\,.$$

Proof. The definition of being smarter than the uniform strategy takes care of the lower bound for optimal arms a: it thus suffices to consider suboptimal arms a. As in the proof of Theorem 2.2.5, we consider a modified bandit problem $\underline{\nu}'$ with $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in \mathcal{D}$ such that $\mu'_a > \mu^\star$, if such a distribution ν'_a exists (otherwise, the first claimed lower bounds equals $-\infty$). From (2.6), we get

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T}\right).$$

We may assume that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]/T \leqslant 1/K$; otherwise, the first claimed bound holds. Since a is the optimal arm under $\underline{\nu}'$ and since the considered strategy is smarter than

the uniform strategy, $\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]/T \geqslant 1/K$. Using that $q \mapsto \mathrm{kl}(p,q)$ is increasing on $[p,1]$, we thus get

$$\mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T}\right) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{1}{K}\right).$$

Lemma 2.5.2 of Appendix 2.5.1 yields

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{1}{K}\right) \geqslant \frac{K}{2}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T} - \frac{1}{K}\right)^2,$$

from which follows, after substitution of the above assumption $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]/T \leqslant 1/K$ in the left-hand side,

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T} \geqslant \frac{1}{K} - \sqrt{\frac{2T}{K^2}\,\mathrm{KL}(\nu_a,\nu'_a)}\,.$$

Taking the supremum of the right-hand side over all $\nu'_a \in \mathcal{D}$ such that $E(\nu'_a) > \mu^\star$ and rearranging concludes the proof. □

2.3.2 Relative lower bound.

Our proof will be based on an assumption of symmetry (milder than requiring that if the arms are permuted in a bandit problem, the algorithm behaves the same way, as in Definition 2.5.6).

Definition 2.3.3. A strategy ψ is pairwise symmetric for optimal arms on $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for each pair of optimal arms $a^\star$ and $a_\star$, the equality $\nu_{a^\star} = \nu_{a_\star}$ entails that, for all $T \geqslant 1$,

$$\big(N_{\psi,a^\star}(T),\ N_{\psi,a_\star}(T)\big) \qquad \text{and} \qquad \big(N_{\psi,a_\star}(T),\ N_{\psi,a^\star}(T)\big)$$

have the same distribution.

Note that the required symmetry is extremely mild as only pairs of *optimal* arms with the *same* distribution are to be considered. What the equality of distributions means is that the strategy should be based only on payoffs and not on the values of the indexes of the arms.

Theorem 2.3.4. *For all models $\mathcal{D}$, for all strategies ψ that are pairwise symmetric for optimal arms on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a and all optimal arms $a^\star$, for all $T \geqslant 1$,*

$$\textit{either}\quad \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{K} \qquad \textit{or}\quad \mathbb{E}_{\underline{\nu}}\left[\frac{\max\big\{N_{\psi,a}(T),\,1\big\}}{\max\big\{N_{\psi,a^\star}(T),\,1\big\}}\right] \geqslant 1 - 2\sqrt{\frac{2T\,\mathrm{KL}(\nu_a,\nu_{a^\star})}{K}}\,.$$

In particular,

$$\forall\, T \leqslant \frac{K}{32\,\mathrm{KL}(\nu_a,\nu_{a^\star})}, \qquad \textit{either}\quad \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{K} \qquad \textit{or}\quad \mathbb{E}_{\underline{\nu}}\left[\frac{\max\big\{N_{\psi,a}(T),\,1\big\}}{\max\big\{N_{\psi,a^\star}(T),\,1\big\}}\right] \geqslant \frac{1}{2}\,.$$

That is, on average, in the small T regime, each suboptimal arm is played at least half the number of times when an optimal arm was played.

Proof. For all arms k, we denote by $N^+_{\psi,k}(T) = \max\big\{N_{\psi,k}(T),\, 1\big\}$. Given a bandit problem $\underline{\nu}$ and a suboptimal arm a, we form an alternative bandit problem $\underline{\nu}'$ given by $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a = \nu_{a^\star}$, where $a^\star$ is an optimal arm of $\underline{\nu}$. In particular, arms a and $a^\star$ are both optimal arms under $\underline{\nu}'$. By the assumption of pairwise symmetry for optimal arms, we have in particular that

$$\mathbb{E}_{\underline{\nu}'}\left[\frac{N^+_{\psi,a}(T)}{N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)}\right] = \mathbb{E}_{\underline{\nu}'}\left[\frac{N^+_{\psi,a^\star}(T)}{N^+_{\psi,a^\star}(T)+N^+_{\psi,a}(T)}\right] = \frac{1}{2}\,.$$

The latter equality and the fundamental inequality (2.6) yield in the present case, through the choice of $Z = N^+_{\psi,a}(T)/\big(N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)\big)$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\ \mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\mathbb{E}_{\underline{\nu}}\left[\frac{N^+_{\psi,a}(T)}{N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)}\right],\ \frac{1}{2}\right). \tag{2.12}$$

The concavity of the function $x \mapsto x/(1+x)$ and Jensen's inequality show that

$$\mathbb{E}_{\underline{\nu}}\left[\frac{N^+_{\psi,a}(T)}{N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)}\right] = \mathbb{E}_{\underline{\nu}}\left[\frac{N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)}{1+N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)}\right] \leqslant \frac{\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big]}{1+\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big]}\,.$$

We can assume that $\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big] \leqslant 1$, otherwise, the result of the theorem is obtained. In this case, the latter upper bound is smaller than $1/2$. Using in addition that $p \mapsto \mathrm{kl}(p,1/2)$ is decreasing on $[0,1/2]$, and assuming that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \leqslant T/K$ (otherwise, the result of the theorem is obtained as well), we get from (2.12)

$$\frac{T}{K}\ \mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big]}{1+\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big]},\ \frac{1}{2}\right).$$

Pinsker's inequality (in its classical form, see Appendix 2.5.1 for a statement) entails the inequality

$$\frac{T}{K}\ \mathrm{KL}(\nu_a,\nu'_a) \geqslant 2\left(\frac{1}{2}-\frac{r}{1+r}\right)^2, \qquad \text{where} \quad r = \mathbb{E}_{\underline{\nu}}\left[\frac{N^+_{\psi,a}(T)}{N^+_{\psi,a^\star}(T)}\right].$$

In particular,

$$\frac{r}{1+r} \geqslant \frac{1}{2} - \sqrt{\frac{T\,\mathrm{KL}(\nu_a,\nu'_a)}{2K}}\,.$$

Applying the increasing function $x \mapsto x/(1-x)$ to both sides, we get

$$r \geqslant \frac{1-\sqrt{2T\,\mathrm{KL}(\nu_a,\nu'_a)/K}}{1+\sqrt{2T\,\mathrm{KL}(\nu_a,\nu'_a)/K}} \geqslant \left(1-\sqrt{\frac{2T\,\mathrm{KL}(\nu_a,\nu'_a)}{K}}\right)^2,$$

where we used $1/(1+x) \geqslant 1-x$ for the last inequality and where we assumed that T is small enough to ensure $1 - \sqrt{2T\,\mathrm{KL}(\nu_a, \nu'_a)/K} \geqslant 0$. Whether this condition is satisfied or not, we have the (possibly void) lower bound

$$r \geqslant 1 - 2\sqrt{\frac{2T\,\mathrm{KL}(\nu_a, \nu'_a)}{K}}\,.$$

The proof is concluded by noting that by definition $\nu'_a = \nu_{a^\star}$. □

2.3.3 Collective lower bound.

In this section, for any given bandit problem $\underline{\nu}$, we denote by $\mathcal{A}^\star(\underline{\nu})$ the set of its optimal arms and by $\mathcal{W}(\underline{\nu})$ the set of its worst arms, i.e., the ones associated with the distributions with the smallest expectation among all distributions for the arms. We also let $A^\star_{\underline{\nu}}$ be the cardinality of $\mathcal{A}^\star(\underline{\nu})$.

We define the following partial order $\preccurlyeq$ on bandit problems: $\underline{\nu}' \preccurlyeq \underline{\nu}$ if

$$\forall a \in \mathcal{A}^\star(\underline{\nu}), \quad \nu_a = \nu'_a \qquad \text{and} \qquad \forall a \notin \mathcal{A}^\star(\underline{\nu}), \quad E(\nu'_a) \leqslant E(\nu_a)\,.$$

In particular, $\mathcal{A}^\star(\underline{\nu}) = \mathcal{A}^\star(\underline{\nu}')$ in this case. The definition models the fact that the bandit problem $\underline{\nu}'$ should be easier than $\underline{\nu}$, as non-optimal arms in $\underline{\nu}'$ are farther away from the optimal arms (in expectation) that in $\underline{\nu}$. Any reasonable strategy should perform better on $\underline{\nu}'$ than on $\underline{\nu}$, which leads to the following definition, where we measure performance in the expected number of times optimal arms are pulled. (Recall that the sets of optimal arms are identical for $\underline{\nu}$ and $\underline{\nu}'$.)

Definition 2.3.5. A strategy ψ is monotonic on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}' \preccurlyeq \underline{\nu}$ in $\mathcal{D}$,

$$\sum_{a^\star \in \mathcal{A}^\star(\underline{\nu}')} \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a^\star}(T)\big] \geqslant \sum_{a^\star \in \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a^\star}(T)\big]\,.$$

Theorem 2.3.6. *For all models $\mathcal{D}$, for all strategies ψ that are pairwise symmetric for optimal arms and monotonic on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, suboptimal arms are collectively sampled at least*

$$\sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant T\left(1 - \frac{A^\star_{\underline{\nu}}}{K} - \frac{A^\star_{\underline{\nu}}\sqrt{2T\,\mathcal{K}^{\max}_{\underline{\nu}}}}{K} - \frac{2A^\star_{\underline{\nu}} T \mathcal{K}^{\max}_{\underline{\nu}}}{K}\right),$$

$$\text{where} \qquad \mathcal{K}^{\max}_{\underline{\nu}} = \min_{w \in \mathcal{W}(\underline{\nu})} \max_{a^\star \in \mathcal{A}^\star(\underline{\nu})} \mathrm{KL}(\nu_w, \nu_{a^\star})\,.$$

In particular,

$$\forall\, T \leqslant \frac{K}{8\,A^\star_{\underline{\nu}}\,\mathcal{K}^{\max}_{\underline{\nu}}}, \qquad \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{2}\left(1 - \frac{A^\star_{\underline{\nu}}}{K}\right).$$

To get a lower bound on the regret from this theorem, we use

$$R_{\psi,\underline{\nu},T} \geqslant \left(\min_{a\notin\mathcal{A}^\star(\underline{\nu})} \Delta_a\right) \sum_{a\notin\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,. \tag{2.13}$$

Proof. We denote by $\widetilde{w}$ some $w \in \mathcal{W}(\underline{\nu})$ achieving the minimum in the defining equation of $\mathcal{K}^{\max}_{\underline{\nu}}$. We construct two bandit models from $\underline{\nu}$. First, the model $\underline{\underline{\nu}}$ differs from $\underline{\nu}$ only at suboptimal arms $a \notin \mathcal{A}^\star(\underline{\nu})$, which we associate with $\underline{\underline{\nu}}_a = \nu_{\widetilde{w}}$. By construction, $\underline{\underline{\nu}} \preccurlyeq \underline{\nu}$.

In the second model $\underline{\underline{\nu}}$, each arm is associated with $\nu_{\widetilde{w}}$, i.e., $\underline{\underline{\nu}}_a = \nu_{\widetilde{w}}$ for all $a \in \{1,\dots,K\}$.

By monotonicity of ψ,

$$\sum_{a\notin\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \sum_{a\notin\mathcal{A}^\star(\underline{\underline{\nu}})} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a}(T)\big]\,.$$

We can therefore focus our attention, for the rest of the proof, on the $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a}(T)\big]$. The strategy is also pairwise symmetric for optimal arms and all arms of $\underline{\underline{\nu}}$ are optimal. This implies in particular that $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,1}(T)\big] = \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a}(T)\big]$ for all arms a, thus $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a}(T)\big] = T/K$ for all arms a.

Now, the bound (2.6) with $Z = \displaystyle\sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \frac{N_{\psi,a^\star}(T)}{T}$ and the bandit models $\underline{\underline{\nu}}$ and $\underline{\underline{\underline{\nu}}}$ gives

$$\begin{aligned}\sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a^\star}(T)\big]\mathrm{KL}(\nu_{\widetilde{w}},\nu_{a^\star}) &\geqslant \mathrm{kl}\left(\sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a^\star}(T)\big]/T, \sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a^\star}(T)\big]/T\right)\\ &= \mathrm{kl}\left(\frac{A^\star_{\underline{\nu}}}{K}, \sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a^\star}(T)\big]/T\right).\end{aligned}$$

By definition of $\mathcal{K}^{\max}_{\underline{\nu}}$ and $\widetilde{w}$, and because $\mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a}(T)\big] = T/K$, we have

$$\sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a^\star}(T)\big]\mathrm{KL}(\nu_{\widetilde{w}},\nu_{a^\star}) \leqslant \frac{T A^\star_{\underline{\nu}} \mathcal{K}^{\max}_{\underline{\nu}}}{K}\,,$$

which yields the inequality

$$\frac{T A^\star_{\underline{\nu}} \mathcal{K}^{\max}_{\underline{\nu}}}{K} \geqslant \mathrm{kl}\left(\frac{A^\star_{\underline{\underline{\nu}}}}{K}, x\right) \qquad \text{where} \qquad x = \frac{1}{T}\sum_{a^\star\in\mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a^\star}(T)\big]\,.$$

We want to upper bound x, in order to get a lower bound on $1-x$. We assume that $x \geqslant A^\star_{\underline{\underline{\nu}}}/K$, otherwise, the bound (2.14) stated below is also satisfied. Pinsker's inequality

(actually, its local refinement stated as Lemma 2.5.2 in Appendix 2.5.1) then ensures that

$$\frac{TA^\star_{\underline{\nu}}\mathcal{K}^{\max}_{\underline{\nu}}}{K} \geqslant \frac{1}{2x}\left(\frac{A^\star_{\underline{\nu}}}{K} - x\right)^2 ,$$

Lemma 2.3.7 below finally entails that

$$x \leqslant \frac{A^\star_{\underline{\nu}}}{K}\left(1 + 2T\mathcal{K}^{\max}_{\underline{\nu}} + \sqrt{2T\mathcal{K}^{\max}_{\underline{\nu}}}\right). \tag{2.14}$$

The proof is concluded by putting all elements together thanks to the monotonicity of ψ and the definition of x:

$$\sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] = T(1-x)\,.$$

□

Lemma 2.3.7. *If $x \in \mathbb{R}$ satisfies $(x-\alpha)^2 \leqslant \beta x$ for some $\alpha \geqslant 0$ and $\beta \geqslant 0$, then $x \leqslant \alpha + \beta + \sqrt{\alpha\beta}$.*

Proof. By assumption, $x^2 - (2\alpha+\beta)x + \alpha^2 \leqslant 0$. We have that x is smaller than the larger root of the associated polynomial, that is,

$$x \leqslant \frac{2\alpha+\beta+\sqrt{(2\alpha+\beta)^2 - 4\alpha^2}}{2} = \frac{2\alpha+\beta+\sqrt{4\alpha\beta+\beta^2}}{2}\,.$$

We conclude with $\sqrt{4\alpha\beta+\beta^2} \leqslant \sqrt{4\alpha\beta} + \sqrt{\beta^2}$. □

2.3.4 Numerical illustrations.

In this section we illustrate some of the bounds stated above for the initial linear regime, namely, the bounds of Theorems 2.3.2 and 2.3.6. It turned out that because of the "or" statement in Theorem 2.3.4, its bound was less easy to illustrate. We need much more difficult bandit problems than the one of Figure 2.1 in order to clearly observe the initial linear phase.

Theorem 2.3.2 is illustrated in Figure 2.2. We observe that in the bandit problems contemplated therein, the expected numbers of pulls of the suboptimal arms considered indeed lie between $T/(2K)$ and T/K in the initial phase, as prescribed by the theorem. We see, however, that this initial phase is probably longer than what was quantified.

Theorem 2.3.6 is illustrated in Figure 2.3. For a large number of arms, the regret lower bound (2.13) deriving as a consequence of the considered theorem is larger than a bound based on the decomposition of the regret (2.1) and Theorem 2.3.2.

2.4 Non-asymptotic bounds for large T.

We restrict our attention to well-behaved models and uniformly super-fast convergent strategies. For a given model $\mathcal{D}$, we denote by $E(\mathcal{D})$ the interior of the set of all expectations of distributions in $\mathcal{D}$. That a model is well-behaved means that the function $\mathcal{K}_{\inf}$

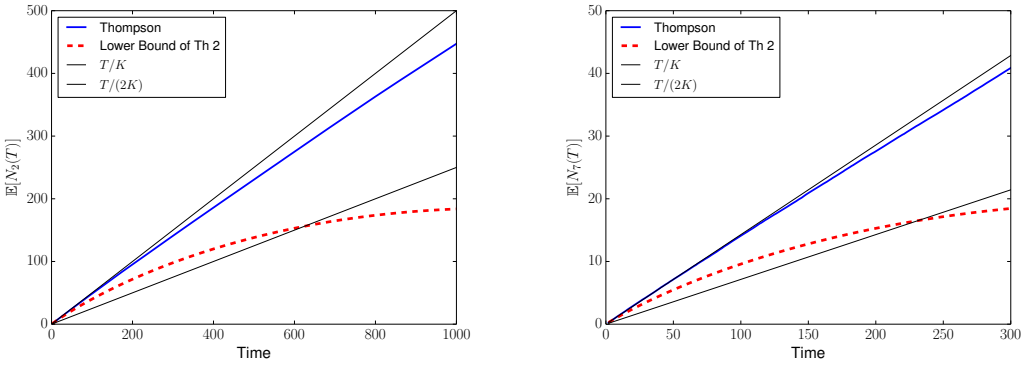

Figure 2.2: Expected number of pulls of the most suboptimal arm for Thompson [1933] Sampling (*blue, solid* line) on Bernoulli bandit problems, versus the lower bound (*red, dashed* line) of Theorem 2.3.2 for the model $\mathcal{D}$ of all Bernoulli distributions; expectations are approximated over $1,000$ runs.

Left: parameters $(\mu_a)_{1 \leqslant a \leqslant 2} = (0.5,\, 0.49)$, with characteristic time $1/\big(8\,\mathcal{K}_{\inf}(\nu_2, \mu^\star, \mathcal{D})\big) \approx 625$.

Right: parameters $(\mu_a)_{1 \leqslant a \leqslant 7} = (0.05,\, 0.048,\, 0.047,\, 0.046,\, 0.045,\, 0.044,\, 0.043)$, with $1/\big(8\,\mathcal{K}_{\inf}(\nu_7, \mu^\star, \mathcal{D})\big) \approx 231$.

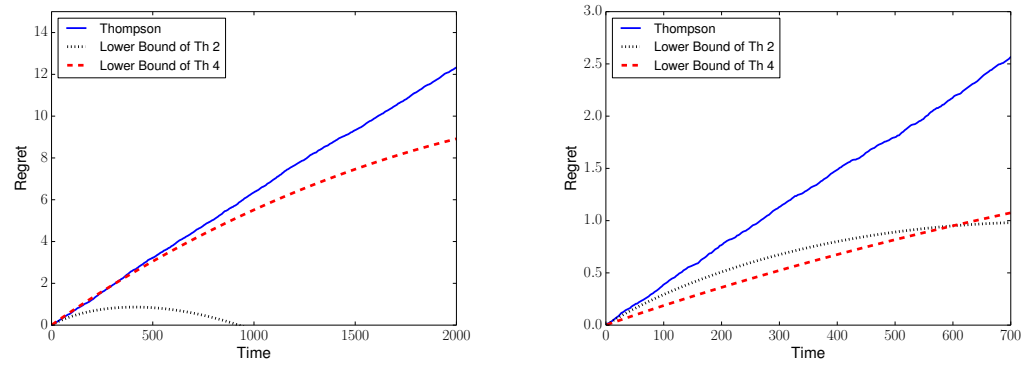

Figure 2.3: Expected regret of Thompson [1933] Sampling (*blue, solid* line) on Bernoulli bandit problems, versus the lower bound (*red, dashed* line) of Theorem 2.3.6 using (2.13) and the lower bound (*black, dotted* line) of Theorem 2.3.2 using (2.1), for the model $\mathcal{D}$ of all Bernoulli distributions; expectations are approximated over $3,000$ runs.

Left: parameters $(\mu_a)_{1 \leqslant a \leqslant 10} = (0.05,\, 0.043,\, \ldots,\, 0.043)$, with characteristic time $K/\big(8\,A^\star_{\underline{\nu}}\,\mathcal{K}^{\max}_{\underline{\nu}}\big) \approx 1,250$.

Right: parameters $(\mu_a)_{1 \leqslant a \leqslant 7} = (0.05,\, 0.048,\, 0.047,\, 0.046,\, 0.045,\, 0.044,\, 0.043)$, with $K/\big(8\,A^\star_{\underline{\nu}}\,\mathcal{K}^{\max}_{\underline{\nu}}\big) \approx 1,619$.

is locally Lipschitz continuous in its second variable, as is made formal in the following definition.

Definition 2.4.1. A model $\mathcal{D}$ is well behaved if there exist two functions $\varepsilon_{\mathcal{D}} : E(\mathcal{D}) \to (0, +\infty)$ and $\omega_{\mathcal{D}} : \mathcal{D} \times E(\mathcal{D}) \to (0, +\infty)$ such that for all distributions $\nu_a \in \mathcal{D}$ and all $x \in E(\mathcal{D})$ with $x > E(\nu_a)$,

$$\forall \varepsilon < \varepsilon_{\mathcal{D}}(x), \qquad \mathcal{K}_{\inf}(\nu_a, x + \varepsilon, \mathcal{D}) \leqslant \mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}) + \varepsilon\, \omega_{\mathcal{D}}(\nu_a, x)\,.$$

We could have considered a more general definition, where the upper bound would have been any vanishing function of ε, not only a linear function of ε. However, all examples considered in this chapter (see Section 2.4.2) can be associated with such a linear difference. Those examples of well-behaved models include parametric families like regular exponential families, as well as more massive classes, like the set of all distributions with bounded support (with or without a constraint on the finiteness of support). Some of these examples, namely, regular exponential families and finitely-supported distributions with common bounded support, were the models studied in Cappé et al. [2013] to get non-asymptotic upper bounds on the regret of the optimal order (2.2).

Definition 2.4.2. A strategy ψ is uniformly super-fast convergent on a model $\mathcal{D}$ if there exists a constant $C_{\psi,\mathcal{D}}$ such that for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a, for all $T \geqslant 2$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \leqslant C_{\psi,\mathcal{D}} \frac{\log T}{\Delta_a^2}\,.$$

Uniform super-fast convergence is a refinement of the notion of uniform fast convergence based on two considerations. First, that there exist such strategies, for instance, the `UCB` strategy of Auer et al. [2002a] on any bounded model $\mathcal{D}$, i.e., a model with distributions all supported within a common bounded interval $[m, M]$. Second, Pinsker's inequality (see Appendix 2.5.1) and Lemma 2.2.3 entail in particular that for such bounded models $\mathcal{D}$,

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}) \geqslant \mathrm{kl}\left(\frac{\mu_a - m}{M - m}, \frac{\mu^\star - m}{M - m}\right) \geqslant \frac{2}{(M-m)^2}\Delta_a^2\,;$$

therefore, the upper bound stated in the definition of uniform super-fast convergence is still weaker than the lower bound (2.2).

Note that Definition 2.4.2 could be relaxed even more: we are mostly interested therein in the logarithmic growth rate $\log T$. We imposed the $C_{\psi,\mathcal{D}}/\Delta_a^2$ upper bound mostly for simplicity and readability of the calculations that lead to Theorem 2.4.3. It would be of course possible to rather consider more abstract problem-dependent constants of the form $C_{\psi,\mathcal{D}}(a, \nu)$, at least as soon as some minimal properties are assumed with respect to the behavior of such constants as functions of the gap $\mu^\star - \mu_a$.

2.4.1 A general non-asymptotic lower bound.

Throughout this subsection, we fix a strategy ψ that is uniformly super-fast convergent with respect to a model $\mathcal{D}$. We recall that we denote by $\mathcal{A}^\star(\underline{\nu})$ the set of optimal arms of the bandit problem $\underline{\nu}$ and let $A^\star_{\underline{\nu}}$ be its cardinality. We adapt the bounds (2.6) and (2.10) by using this time

$$Z = \frac{1}{T} \sum_{a^\star \in \mathcal{A}^\star(\underline{\nu})} N_{\psi,a^\star}(T)$$

and $\mathrm{kl}(p,q) \geqslant p \log(1/q) - \log 2$, see (2.11). For all bandit problems $\underline{\nu}'$ that only differ from $\underline{\nu}$ as far a suboptimal arm a is concerned, whose distribution of payoffs $\nu'_a \in \mathcal{D}$ is such that $\mu'_a = E(\nu'_a) > \mu^\star$, we get

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)} \left(\mathbb{E}_{\underline{\nu}}[Z] \, \log \frac{1}{\mathbb{E}_{\underline{\nu}'}[Z]} - \log 2 \right). \tag{2.15}$$

We restrict our attention to distributions $\nu'_a \in \mathcal{D}$ such that the gaps for $\underline{\nu}'$ associated with optimal arms $a^\star \in \mathcal{A}^\star(\underline{\nu})$ of $\underline{\nu}$ satisfy $\underline{\Delta} = \mu'_a - \mu^\star \geqslant \varepsilon$, for some parameter $\varepsilon > 0$ to be defined by the analysis. By uniform super-fast convergence, on the one hand,

$$\mathbb{E}_{\underline{\nu}}[Z] = 1 - \frac{1}{T} \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant 1 - \frac{1}{T} \left(C_{\psi,\mathcal{D}} \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \log T \right);$$

on the other hand,

$$\mathbb{E}_{\underline{\nu}'}[Z] = \frac{1}{T} \sum_{a^\star \in \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big] \leqslant \frac{A^\star_{\underline{\nu}} \, C_{\psi,\mathcal{D}}}{\underline{\Delta}^2} \frac{\log T}{T}.$$

Denoting

$$H(\underline{\nu}) = \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \tag{2.16}$$

and using that $\underline{\Delta} \geqslant \varepsilon$, a substitution of the two super-fast convergence inequalities into (2.15) and an optimization over the considered distributions ν'_a leads to

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})} \left(1 - C_{\psi,\mathcal{D}} H(\underline{\nu}) \frac{\log T}{T}\right) \log \frac{T\varepsilon^2}{A^\star_{\underline{\nu}} \, C_{\psi,\mathcal{D}} \log T} - \frac{\log 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})}. \tag{2.17}$$

The obtained bound holds for all $T \geqslant 2$ (as in the definition of uniform super-fast convergence); however, for small values of T, it might be negative, thus useless.

To proceed, we use the fact that the model $\mathcal{D}$ is well-behaved to relate $\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})$ to $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})$. Since $1/(1+x) \geqslant 1 - x$ for all $x \geqslant 0$, we get by Definition 2.4.1

$$\forall \varepsilon < \varepsilon_{\mathcal{D}}(\mu^\star), \qquad \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})} \left(1 - \varepsilon \, \frac{\omega_{\mathcal{D}}(\nu_a, \mu^\star)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\right).$$

Now, we set $\varepsilon = \varepsilon_T = (\log T)^{-4}$. Many other choices would have been possible, but this one is such that $\varepsilon_T \leqslant 0.0005$ already for $T \geqslant 1\,000$. Putting all things together, from (2.17), from the fact that $(1-a)(1-b)(1-c) \geqslant 1-(a+b+c)$ when $0 \leqslant a, b, c \leqslant 1$, and from the bound $A^\star_{\underline{\nu}} \leqslant K$, we get the following theorem.

Theorem 2.4.3. *For all uniformly super-fast convergent strategies ψ on well-behaved models $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a,*

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\big(1 - (a_T + b_T + c_T)\big) - \frac{\log 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\,, \tag{2.18}$$

for all $T \geqslant 2$ large enough so that $(\log T)^{-4} < \varepsilon_{\mathcal{D}}(\mu^\star)$ and

$$a_T = \frac{\omega_{\mathcal{D}}(\nu_a, \mu^\star)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}(\log T)^{-4}\,, \qquad b_T = C_{\psi,\mathcal{D}} H(\underline{\nu}) \frac{\log T}{T}\,, \qquad c_T = \frac{\log\big(K\, C_{\psi,\mathcal{D}} (\log T)^9\big)}{\log T}\,,$$

are all smaller than 1, where $H(\underline{\nu})$ was defined in (2.16).

Remark 2.4.4. We have $(a_T + b_T + c_T)\log T = O\big(\log(\log T)\big)$. The non-asymptotic bound (2.18) is therefore of the form

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})} - O\big(\log(\log T)\big)\,.$$

Note that the second-order term of typical non-asymptotic upper bounds (e.g., by [Cappé et al., 2013]) had long been of the form $+(\log T)^\alpha$ for some $\alpha \in (0,1)$. But recently, Honda and Takemura [2015, Theorem 5] showed that at least for models containing distributions that have each a bounded support, the second-order is of order $-\log(\log T)$. Our lower bound above thus shows the optimality of the order of magnitude of this second-order term.

2.4.2 Two (and a half) examples of well-behaved models.

We consider first distributions with common bounded support (and the subclass of such distributions with finite support); and then, regular exponential families. The latter and the subclass of distributions with finite and bounded support are the two models for which Cappé et al. [2013] could prove non-asymptotic upper bounds matching the lower bound (2.2).

Distributions with common bounded support. We denote by $\mathcal{M}\big([0,M]\big)$ the set of all probability distributions over $[0,M]$, equipped with its Borel σ–algebra, and restrict our model to such distributions with expectation not equal to M.

Lemma 2.4.5. *In the model $\mathcal{D} = \Big\{\mathrm{m} \in \mathcal{M}\big([0,M]\big) : E(\mathrm{m}) < M\Big\}$, we have*

$$\forall \mathrm{m} \in \mathcal{D}, \quad \forall \mu^\star \in [0,M), \quad \forall \varepsilon \in \big(0, (M-\mu^\star)/2\big),$$
$$\mathcal{K}_{\inf}(\mathrm{m}, \mu^\star + \varepsilon, \mathcal{D}) \leqslant \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star, \mathcal{D}) - \log\left(1 - \frac{2\varepsilon}{M-\mu^\star}\right).$$

In particular, for all $\mathrm{m} \in \mathcal{D}$ *and* $\mu^\star \in [0, M)$,

$$\forall \varepsilon \in \big(0, (M-\mu^\star)/4\big), \qquad \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star + \varepsilon, \mathcal{D}) \leqslant \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star, \mathcal{D}) + \frac{4\varepsilon}{M-\mu^\star}\,.$$

Proof. We fix m, $\mu^\star$ and ε as indicated for the first bound; in particular, $\mu^\star + \varepsilon < M$. Since m is a probability distribution, it has at most countably many atoms; therefore, there exists some $x \in (\mu^\star + \varepsilon, M)$ such that $\mathrm{m}(\{x\}) = 0$ and $x \geqslant (M + \mu^\star)/2$. In particular, m and the Dirac measure δ_x at this point are singular measures.

We consider some $\mathrm{m}' \in \mathcal{D}$ such that $E(\mathrm{m}') > \mu^\star$ and $\mathrm{m} \ll \mathrm{m}'$ (i.e., m is absolutely continuous with respect to m'). Such distributions exist and they are the only interesting ones in the defining infimum of $\mathcal{K}_{\inf}(\mathrm{m}, \mu^\star, \mathcal{D})$. We associate with m' the distribution

$$\mathrm{m}'_\alpha = (1-\alpha)\mathrm{m}' + \alpha\delta_x\,, \qquad \text{for the value} \qquad \alpha = \frac{\varepsilon}{x-\mu^\star} \in (0,1)\,.$$

The expectation of m'_α satisfies

$$E\big(\mathrm{m}'_\alpha\big) > (1-\alpha)\mu^\star + \alpha x = \mu^\star + \alpha(x-\mu^\star) = \mu^\star + \varepsilon\,. \tag{2.19}$$

Now, $\mathrm{m} \ll \mathrm{m}'$ entails that $\mathrm{m} \ll \mathrm{m}'_\alpha$ as well, with respective densities satisfying (because m and δ_x are singular)

$$\frac{\mathrm{dm}}{\mathrm{dm}'_\alpha} = \frac{1}{1-\alpha}\,\frac{\mathrm{dm}}{\mathrm{dm}'} \qquad \text{and} \qquad \frac{\mathrm{dm}}{\mathrm{dm}'_\alpha}(x) = 0\,.$$

Therefore,

$$\mathrm{KL}(\mathrm{m}, \mathrm{m}'_\alpha) = \int \left(\log \frac{\mathrm{dm}}{\mathrm{dm}'_\alpha}\right) \mathrm{dm} = \log\frac{1}{1-\alpha} + \int \left(\log \frac{\mathrm{dm}}{\mathrm{dm}'}\right) \mathrm{dm} = \log\frac{1}{1-\alpha} + \mathrm{KL}(\mathrm{m}, \mathrm{m}')\,.$$

Since α decreases with x and $x \geqslant (M+\mu^\star)/2$, we get $\alpha \leqslant 2\varepsilon/(M-\mu^\star)$. We substitute this bound in the inequality above and take the infimum in both sides, considering (2.19), to get the first claimed bound. The second bound follows from the inequality $-\log(1-x) \leqslant 2x$ for $x \in [0, 1/2]$. □

Remark 2.4.6. We denote by $\mathcal{M}_{\mathrm{fin}}\big([0,M]\big)$ the subset of $\mathcal{M}\big([0,M]\big)$ formed by probability distributions with finite support. The proof above shows that the bound of Lemma 2.4.5 also holds for the model

$$\mathcal{D} = \Big\{\mathrm{m} \in \mathcal{M}_{\mathrm{fin}}\big([0,M]\big) : E(\mathrm{m}) < M\Big\}\,.$$

Regular exponential families. Another example of well-behaved models is given by regular exponential families, see Lehmann and Casella [1998] for a thorough exposition or Cappé et al. [2013] for an alternative exposition focused on multi-armed bandit problems.

Such a family $\mathcal{D}$ is indexed by an open set $I = (m, M)$, where for each $\mu \in I$ there exists a unique distribution $\nu_\mu \in \mathcal{D}$ with expectation μ. (The bounds m and M can be equal to $\pm\infty$.) A key property of such a family is that the Kullback-Leibler divergence between two of its elements can be represented[1] by a twice differentiable and strictly convex function $g : I \to \mathbb{R}$, with increasing first derivative $\dot{g}$ and continuous second derivative $\ddot{g} \geqslant 0$, in the sense that

$$\forall\, (\mu, \mu') \in I^2, \qquad \mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = g(\mu) - g(\mu') - (\mu - \mu')\, \dot{g}(\mu')\,. \tag{2.20}$$

In particular, $\mu' \mapsto \mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big)$ is strictly convex on I, thus is increasing on $[\mu, M)$. This entails that

$$\forall\, (\mu, \mu^\star) \in I^2 \;\; \text{s.t} \;\; \mu < \mu^\star, \qquad \mathcal{K}_{\inf}(\nu_\mu, \mu^\star, \mathcal{D}) = \mathrm{KL}\big(\nu_\mu, \nu_{\mu^\star}\big)\,. \tag{2.21}$$

In the lemma below, we restrict our attention to $\varepsilon > 0$ such that $\mu^\star + \varepsilon \in I$, e.g., to $\varepsilon < B_{\mu^\star}$ where

$$B_{\mu^\star} = \min\left\{ \frac{M - \mu^\star}{2},\, 1 \right\}. \tag{2.22}$$

The minimum with 1 is considered merely for $B_{\mu^\star}$ to always have a finite value; otherwise, the bound in the lemma below would be uninformative.

Lemma 2.4.7. *In a model $\mathcal{D}$ given by a regular exponential family indexed by $I = (m, M)$ and whose Kullback-Leibler divergence (2.20) is represented by a function g, we have, with the notation (2.22),*

$$\forall\, \mu < \mu^\star \text{ of } I, \quad \forall\, 0 < \varepsilon < B_{\mu^\star}, \qquad \mathcal{K}_{\inf}(\nu_\mu, \mu^\star{+}\varepsilon, \mathcal{D}) \leqslant \mathcal{K}_{\inf}(\nu_\mu, \mu^\star, \mathcal{D}) {+} \varepsilon \left(\mu^\star {+} B_{\mu^\star} {-} \mu\right) G_{\mu^\star}$$

where $G_{\mu^\star} = \max\big\{\ddot{g}(x) : \quad \mu^\star \leqslant x \leqslant \mu^\star + B_{\mu^\star}\big\}$.

Proof. Since $\mu < \mu^\star$, we get by (2.20) and (2.21)

$$\begin{aligned}
&\mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon, \mathcal{D}) - \mathcal{K}_{\inf}(\nu_\mu, \mu^\star, \mathcal{D}) \\
&\quad = \;\; g(\mu^\star) - g(\mu^\star + \varepsilon) - \big(\mu - (\mu^\star + \varepsilon)\big)\, \dot{g}(\mu^\star + \varepsilon) + (\mu - \mu^\star)\, \dot{g}(\mu^\star) \\
&\quad = \;\; \underbrace{g(\mu^\star) - g(\mu^\star + \varepsilon) + \varepsilon\, \dot{g}(\mu^\star)}_{\leqslant 0} + \big((\mu^\star + \varepsilon) - \mu\big)\big(\dot{g}(\mu^\star + \varepsilon) - \dot{g}(\mu^\star)\big)\,,
\end{aligned}$$

where the inequality is obtained by convexity of g. The proof is concluded by an application of the mean-value theorem,

$$\dot{g}(\mu^\star + \varepsilon) - \dot{g}(\mu^\star) \leqslant \varepsilon \max_{(\mu^\star, \mu^\star + \varepsilon)} \ddot{g}\,,$$

and the bound $\varepsilon \leqslant B_{\mu^\star}$. □

[1] This function g has an intrinsic definition as the convex conjugate of the log-normalization function b in the natural parameter space Θ, where b can also be seen as a primitive of the expectation function $\Theta \to I$. But these properties are unimportant here.

The upper bound obtained on $\mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon, \mathcal{D}) - \mathcal{K}_{\inf}(\nu_\mu, \mu^\star, \mathcal{D})$ equals $\varepsilon\big(\mu^\star + B_{\mu^\star} - \mu\big) G_{\mu^\star}$. The examples below propose concrete upper bounds for $G_{\mu^\star}$ in different exponential families. None of these upper bounds actually involves $B_{\mu^\star}$ as various monotonicity arguments can be invoked.

Example 2.4.8. For Poisson distributions, we have $I = (0, +\infty)$ and

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \mu' - \mu + \mu \log \frac{\mu}{\mu'} \,.$$

We may take $g(\mu) = \mu \log \mu - \mu$, so that $\ddot{g}(\mu) = 1/\mu$ and $G_{\mu^\star} = 1/\mu^\star$.

Example 2.4.9. For Gamma distributions with known shape parameter $\alpha > 0$ (e.g., the exponential distributions when $\alpha = 1$), we have $I = (0, +\infty)$ and

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \alpha \left(\frac{\mu}{\mu'} - 1 - \log \frac{\mu}{\mu'} \right) .$$

We may take $g(\mu) = -\alpha \log \mu$, so that $\ddot{g}(\mu) = \alpha/\mu^2$ and $G_{\mu^\star} = \alpha/(\mu^\star)^2$.

Example 2.4.10. For Gaussian distributions with known variance $\sigma^2 > 0$, we have $I = (0, +\infty)$ and

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \frac{(\mu - \mu')^2}{2\sigma^2} \,.$$

We may take $g(\mu) = \mu^2/(2\sigma^2)$, so that $\ddot{g}(\mu) = 1/\sigma^2$ and $G_{\mu^\star} = 1/\sigma^2$.

Example 2.4.11. For binomial distributions for n samples (e.g., Bernoulli distributions when $n = 1$), we have $I = (0, n)$ and

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \mu \log \frac{\mu}{\mu'} + (n - \mu) \log \frac{n - \mu}{n - \mu'} \,.$$

We may take $g(\mu) = \mu \log \mu + (n-\mu) \log(n-\mu)$, so that $\ddot{g}(\mu) = n/\big(\mu(n-\mu)\big)$. A possible upper bound is

$$G_{\mu^\star} \leqslant \frac{2n}{\mu^\star(n - \mu^\star)} \,.$$

This can be seen by noting that $B_{\mu^\star} \leqslant (n - \mu^\star)/2$ so that any $\mu \in [\mu^\star, \mu^\star + B_{\mu^\star}]$ is such that $\mu \geqslant \mu^\star$ and $n - \mu \geqslant n - \mu^\star - B_{\mu^\star} \geqslant (n - \mu^\star)/2$.

2.5 Elements of Proofs

2.5.1 Reminder of some elements of information theory.

For the sake of self-completeness we recall two selected basic facts pertaining to Kullback-Leibler divergences.

The data-processing inequality. The most elegant proof we are aware of relies on a conditional Jensen's inequality applied to $t \mapsto t \log t$; see Ali and Silvey [1966b] or the proof of Lemma 6.8.6.

Lemma 2.5.1. *Consider a measurable space $(\Gamma, \mathcal{G})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, any other $(\Gamma', \mathcal{G}')$ measurable space, and any random variable $X : (\Gamma, \mathcal{G}) \to (\Gamma', \mathcal{G}')$. Then,*

$$\mathrm{KL}\big(\mathbb{P}_1^X, \mathbb{P}_2^X\big) \leqslant \mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2)\,,$$

where $\mathbb{P}_1^X$ and $\mathbb{P}_2^X$ denote the respective distributions of X under $\mathbb{P}_1$ and $\mathbb{P}_2$.

On local refinements of Pinsker's inequality. Pinsker's inequality reads, for Bernoulli distributions, in its most classical form:

$$\forall (p,q) \in [0,1]^2, \qquad \mathrm{kl}(p,q) \geqslant 2(p-q)^2\,. \tag{2.23}$$

The lemma below offers a local refinement of Pinsker's inequality for Bernoulli distributions; the classical form (2.23) follows by noting that $x(1-x) \leqslant 1/4$ for $x \in [0,1]$. Cappé et al. [2013, Lemma 3 in Appendix A.2.1] offer an extension of this local refinement to any one-parameter regular exponential family.

Lemma 2.5.2. *For $0 \leqslant p < q \leqslant 1$, we have*

$$\mathrm{kl}(p,q) \geqslant \frac{1}{2 \max\limits_{x \in [p,q]} x(1-x)} (p-q)^2 \geqslant \frac{1}{2q}(p-q)^2\,.$$

Proof. We may assume that $p > 0$ and $q < 1$, since for $p = 0$, the result follows by continuity, and for $q = 1$, the inequality is void, as $\mathrm{kl}(p,1) = +\infty$ when $p < 1$. The first and second derivative of kl equal

$$\frac{\partial}{\partial p}\mathrm{kl}(p,q) = \log p - \log(1-p) - \log q + \log(1-q) \quad \text{and} \quad \frac{\partial^2}{\partial^2 p}\mathrm{kl}(p,q) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}\,.$$

By Taylor's equality, there exists $r \in [p,q]$ such that

$$\mathrm{kl}(p,q) = \underbrace{\mathrm{kl}(q,q)}_{=0} + (p-q)\underbrace{\frac{\partial}{\partial p}\mathrm{kl}(q,q)}_{=0} + \frac{(p-q)^2}{2}\underbrace{\frac{\partial^2}{\partial^2 p}\mathrm{kl}(r,q)}_{=1/(r(1-r))}\,.$$

The proof of the first inequality is concluded by upper bounding $r(1-r)$ by $\max\limits_{x \in [p,q]} x(1-x)$.

The second inequality follows from $\max\limits_{x \in [p,q]} x(1-x) \leqslant \max\limits_{x \in [p,q]} x \leqslant q$. □

2.5.2 Re-derivation of other earlier lower bounds

In this section, we re-derive the bounds discussed in Section 2.1.3, based on our fundamental inequality (2.6). We do so to illustrate the power and the versatility of (2.6). However, we point out again that the lower bounds discussed here are much weaker than the ones derived in the main body of the chapter: in the terminology of Section 2.1.3, they are of the form (well-chosen) rather than of the form (all).

Distribution-free lower bound.

We consider the bound (2.3) recalled in Section 2.1.3. More specifically, we re-prove Theorem A.2 of Auer et al. [2002b], from which the stated bound (2.3) follows by optimization over ε.

Theorem 2.5.3. *Consider the bandit model $\mathcal{D} = \mathcal{M}\big([0,1]\big)$ of all probability distributions over $[0,1]$. For all $\varepsilon \in (0,1/2)$, for all strategies ψ, there exists a bandit problem $\underline{\nu}'$ in $\mathcal{M}\big([0,1]\big)$ such that*

$$R_{\psi,\underline{\nu}',T} \geqslant T\varepsilon \left(1 - \frac{1}{K} - \frac{1}{2}\sqrt{\frac{T}{K}\log\frac{1}{1-4\varepsilon^2}}\right).$$

This problem $\underline{\nu}'$ can be given by Bernoulli distributions, with parameters $1/2$ for all arms but one, for which the parameter is $1/2+\varepsilon$.

As a consequence, the worst-case regret of any strategy ψ against all bandit problems $\underline{\nu}$ in $\mathcal{M}\big([0,1]\big)$ is lower bounded as announced in (2.3):

$$\sup_{\underline{\nu}} R_{\psi,\underline{\nu},T} \geqslant \sup_{\varepsilon\in(0,1/2)} T\varepsilon \left(1 - \frac{1}{K} - \frac{1}{2}\sqrt{\frac{T}{K}\log\frac{1}{1-4\varepsilon^2}}\right) \geqslant \frac{1}{20}\min\Big\{\sqrt{KT},\, T\Big\}.$$

The second inequality above is proved by a simple calculation indicated after the proof of Theorem A.2 of Auer et al. [2002b]: pick $\varepsilon = \min\big\{\sqrt{K/T},\, 1\big\}/4$ and use $-\log(1-u) \leqslant \big(4\log(4/3)\big)u$ for $u \in (0,1/4)$. The constant $1/20$ can actually be improved into $1/8$, see Cesa-Bianchi and Lugosi [2006, Theorem 6.11].

Proof. We fix a strategy and $\varepsilon \in (0,1/2)$. We denote by $\underline{\nu}$ the bandit problem where all distributions are given by Bernoulli distributions with parameter $1/2$. There exists an arm $k \in \{1,\dots,K\}$ such that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big] \leqslant T/K$, as these K numbers of pulls sum up to T. We define the bandit problem $\underline{\nu}'$ by $\nu'_a = \nu_a$ for $a \neq k$, that is, ν'_a is a symmetric Bernoulli distribution, while ν'_k is the Bernoulli distribution with parameter $1/2+\varepsilon$. By (2.1), we have

$$R_{\psi,\underline{\nu}',T} = \sum_{a\neq k} \varepsilon\, \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big] = T\varepsilon\left(1 - \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right). \tag{2.24}$$

A direct computation of $\mathrm{kl}(1/2,\, 1/2+\varepsilon)$ and the application of (2.6) indicate that

$$\begin{aligned}\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{2}\log\frac{1}{1-4\varepsilon^2} &= \mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]\,\mathrm{kl}\left(\frac{1}{2},\,\frac{1}{2}+\varepsilon\right)\\ &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T},\,\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right).\end{aligned}$$

Now, Pinsker's inequality (in its classical form, see Appendix 2.5.1) ensures that

$$\begin{aligned}\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{2}\log\frac{1}{1-4\varepsilon^2} &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T},\,\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right)\\ &\geqslant 2\left(\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}-\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T}\right)^2.\end{aligned}$$

Solving for $\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]/T$, based on whether $\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]/T$ is larger or smaller than $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]/T$, we get, in all cases,

$$\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T} \leqslant \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T}+\frac{1}{2}\sqrt{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]\log\frac{1}{1-4\varepsilon^2}}\,.$$

The proof is concluded by substituting the fact that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big] \leqslant T/K$ by definition of k, and by combining the obtained inequality with (2.24). □

The short proof above actually re-uses absolutely all the original arguments of Auer et al. [2002b]: the same Bernoulli distributions, the chain rule for Kullback-Leibler divergences, Pinsker's inequality. It is merely stated in a compact way, that puts under the same umbrella the distribution-dependent and the distribution-free lower bounds for multi-armed bandit problems.

Following the same lines of the proof of Theorem 2.5.3 one can prove the same type of theorem for the family of bandit problems $\mathcal{F}$ described in Section 3.2.

Theorem 2.5.4. *Consider the bandit model $\mathcal{D}=\mathcal{F}$ of the Section 3.2. For all $\mu\in[\mu^-,\mu^+)$, for all $\varepsilon\in(0,\mu^+-\mu)$, for all strategies ψ, there exists a bandit problem $\underline{\nu}'$ in $\mathcal{F}$ such that*

$$R_{\psi,\underline{\nu}',T} \geqslant T\varepsilon\left(1-\frac{1}{K}-\frac{1}{2}\sqrt{\frac{T}{K}d(\mu,\mu+\varepsilon)}\right).$$

This problem $\underline{\nu}'$ can be given by distributions $\nu_{b'^{-1}(.)}$ with parameters μ for all arms but one, for which the parameter is $\mu+\varepsilon$.

It remains to control the divergence $d(\mu,\mu+\varepsilon)$. Let $V' := \sup_{\mu\in[\mu^-,\mu^+]} b''\big(b'^{-1}(\mu)\big)$ be the maximum of the variance in the one parameter exponential family over the interval $[\mu^-,\mu^+]$. By the continuity of the variance there exists (μ_0,ε_0) such that $[\mu_0,\mu_0+\varepsilon_0]\subset[\mu^-,\mu^+]$ and for all $\mu\in[\mu_0,\mu_0+\varepsilon_0]$ it holds

$$b''\big(b'^{-1}(\mu)\big) \geqslant \frac{V'}{2}\,.$$

Thus, thanks to a Taylor expansion one obtains for all $\varepsilon \in (0, \varepsilon_0]$,

$$d(\mu_0, \mu_0 + \varepsilon) = \int_{\mu_0}^{\mu_0+\varepsilon} \frac{1}{b''\big(b'^{-1}(\mu)\big)}(x - \mu_0)\mathrm{d}x \leqslant \frac{1}{V'}\varepsilon^2 \,.$$

As above, the Theorem 2.5.4 entails that with the choice $\varepsilon = \min\{\sqrt{V'K/T}, \varepsilon_0\}/4$,

$$\sup_{\underline{\nu}} R_{\psi,\underline{\nu},T} \geqslant T\varepsilon\left(1 - \frac{1}{K} - \frac{1}{2}\sqrt{\frac{T}{V'K}\varepsilon^2}\right) \geqslant \frac{1}{16}\min\Big\{\sqrt{V'KT},\, \varepsilon_0^2 T\Big\} \,. \tag{2.25}$$

2.5.3 Lower bounds for the case when $\mu^\star$ or the gaps Δ are known.

We consider here the second framework discussed in Section 2.1.3, with sub-Gaussian bandit problems. For simplicity, and following Bubeck et al. [2013a], we restrict our attention to lower bounds for two-armed bandit problems (i.e., for $K = 2$).

Known largest expected payoff $\mu^\star$ but unknown gap Δ. The lower bound stated in Theorem 2.5.5 below corresponds to Theorem 8 of Bubeck et al. [2013a], later revisited by the authors, see Bubeck et al. [2013b]. It turns out that, as hinted at in, e.g., Faure et al. [2015, end of Section 1.4], the initially claimed $\log T$ dependency is incorrect and a bounded regret can be guaranteed. As shown in Theorem 2.6.1 in the next section, this bound on the regret can be as small as $\log(1/\Delta)/\Delta$. The lower bound we could get using our techniques is of order $1/\Delta$.

To state it, we restrict our attention to strategies ψ symmetric in some sense, e.g., in the sense of Definition 2.3.3 stated later on. We actually need very little symmetry here: the considered strategies ψ should just be such that in the bandit problem $\underline{\nu}_0 = \big(\mathcal{N}(0,1), \mathcal{N}(0,1)\big)$, in which the two arms have the same distribution,

$$\mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,1}(T)\big] = \mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,2}(T)\big] = \frac{T}{2} \,. \tag{2.26}$$

Of course, all reasonable strategies are usually even more symmetric than that: they are usually stable by permutations over the arms (i.e., they base their decisions only on the payoffs received, not on the labeling of the arms).

Theorem 2.5.5. *For all $\Delta > 0$ we consider $\underline{\nu}_\Delta = \big(\mathcal{N}(0,1), \mathcal{N}(-\Delta,1)\big)$ and $\underline{\nu}_0 = \big(\mathcal{N}(0,1), \mathcal{N}(0,1)\big)$. For all strategies ψ that are symmetric in the sense of (2.26), for all $\Delta > 0$, for all $T \geqslant 1$,*

$$\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant \frac{1}{\Delta^2 + 1/T} \qquad \textit{and} \qquad R_{\psi,\underline{\nu}_\Delta,T} \geqslant \frac{\Delta}{\Delta^2 + 1/T} \,.$$

In addition, for all strategies ψ and for all T such that $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 1$,

$$\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant \min\left\{\frac{2\log 2}{\Delta^2 + 2\log(4T)/T},\, \frac{T}{2}\right\} \qquad \textit{and}$$
$$R_{\psi,\underline{\nu}_\Delta,T} \geqslant \min\left\{\frac{2(\log 2)\Delta}{\Delta^2 + 2\log(4T)/T},\, \frac{T\Delta}{2}\right\} .$$

Note that the constraint that $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 1$ is satisfied for all $T \geqslant K$ by most of the reasonable strategies, as the latter typically start by playing each arm once (in a random order).

Proof. We first note that $R_{\psi,\underline{\nu}_\Delta,T} = \Delta\ \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]$. Inequality (2.6) entails that

$$\begin{aligned}\frac{\Delta^2}{2}\,\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] &= \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]\ \mathrm{KL}\big(\mathcal{N}(-\Delta,1),\,\mathcal{N}(0,1)\big)\\ &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T},\ \frac{1}{2}\right), \qquad (2.27)\end{aligned}$$

where we used respectively, for the two equalities, the closed-form expression for the Kullback-Leibler divergences between Gaussian distribution with the same variance and the symmetry assumption on the strategy. Pinsker's inequality (in its classical form, see Appendix 2.5.1), followed by the inequality

$$\forall\, x \in \mathbb{R}, \qquad 2\left(\frac{1}{2} - x\right)^2 \geqslant \frac{1}{2} - 2x\,,$$

yields

$$\frac{\Delta^2}{2}\,\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 2\left(\frac{1}{2} - \frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T}\right)^2 \geqslant \frac{1}{2} - 2\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T}\,.$$

Simple manipulations entail the first claimed bound on $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]$.

For the second one, given the form of the lower bound, which involves a minimum with $T/2$, it suffices to consider the case when $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]\,/T \leqslant 1/2$. We use that

$$\mathrm{kl}(x,\,1/2) = \log 2 - h(x)\,, \qquad \text{where} \qquad h(x) = -\big(x\log x + (1-x)\log(1-x)\big)$$

is the binary entropy function. Now, Calabro [2009, page 8] indicates that $h(x) \leqslant x\log(4/x)$ for all $x \in [0, 1/2]$, so that, restricting our attention to $x \geqslant 1/T$, we get

$$\forall\, x \in [1/T,\,1/2], \qquad \mathrm{kl}\left(x,\,\frac{1}{2}\right) \geqslant \log 2 - x\log\left(\frac{4}{x}\right) \geqslant \log 2 - x\log(4T)\,.$$

Substituting this inequality into (2.27), using that $x = \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]\,/T$ lies in $[1/T,\,1/2]$, concludes the proof. □

The proof above, which is simple and direct, illustrates the interest of Inequality (2.6) over the standard approaches used so far to prove lower bounds in the same or similar settings.

Known gap Δ but unknown largest expected payoff $\mu^\star$. The lower bound stated in Theorem 2.5.7 below corresponds to Theorem 6 of Bubeck et al. [2013a]. It shows the optimality of the performance bound $\log(T\Delta^2)/\Delta$ on the regret of the Improved–UCB strategy introduced by Auer and Ortner [2010] and further studied by Garivier et al. [2016]. The latter improved the constant in the leading term, which equals $\log(T\Delta^2)/(2\Delta)$ when the gap Δ between the expected payoffs between the two Gaussian arms with variance 1 is known.

We denote by W the Lambert function: for all $u \geqslant 0$, there exists a unique $v \geqslant 0$ such that $u\exp(u) = v$, which is denoted by $v = W(u)$. The Lambert function W is increasing on $[0, +\infty)$. One may easily check that

$$\forall x \geqslant \mathrm{e}, \qquad \log(x) - \log\bigl(\log(x)\bigr) \leqslant W(x) \leqslant \log(x)\,.$$

We state below two lower bounds: one for all strategies ψ, in terms of a maximum between two regrets; and one for strategies that are symmetric and invariant by translation. These properties of symmetry and invariance by translation are most natural requirements. To define them, for all $c \in \mathbb{R}$ and all distributions ν, we denote by $\tau_c(\nu)$ the distribution of $Y + c$ when $Y \sim \nu$.

Definition 2.5.6. A strategy ψ for K–armed bandits is symmetric and invariant by translation of the payoffs if for all permutations σ of $\{1, \dots, K\}$, all $c \in \mathbb{R}$, and all $T \geqslant 1$, the distribution of $\bigl(N_{\psi,1}(T), \dots, N_{\psi,K}(T)\bigr)$ in the bandit problem $(\nu_1, \dots, \nu_K)$ is equal to the one of $\bigl(N_{\psi,\sigma^{-1}(1)}(T), \dots, N_{\psi,\sigma^{-1}(K)}(T)\bigr)$ in the bandit problem $\bigl(\tau_c(\nu_{\sigma(1)}), \dots, \tau_c(\nu_{\sigma(K)})\bigr)$.

Theorem 2.5.7. *We fix $\Delta > 0$ and consider $\underline{\nu}_1 = \bigl(\mathcal{N}(0,1), \mathcal{N}(-\Delta,1)\bigr)$ and $\underline{\nu}_2 = \bigl(\mathcal{N}(0,1), \mathcal{N}(\Delta,1)\bigr)$. Then, for all strategies ψ, for all $T \geqslant 1$,*

$$\max\bigl\{R_{\psi,\underline{\nu}_1,T},\, R_{\psi,\underline{\nu}_2,T}\bigr\} \geqslant \min\left\{\frac{W\bigl(T\Delta^2/1.2\bigr)}{2\Delta},\, \frac{T\Delta}{2}\right\}. \tag{2.28}$$

Or, alternatively, for all strategies ψ that are symmetric and invariant by translation of the payoffs, for all $T \geqslant 1$,

$$R_{\psi,\underline{\nu}_1,T} = R_{\psi,\underline{\nu}_2,T} \geqslant \frac{W\bigl(T\Delta^2/1.2\bigr)}{2\Delta}\,.$$

Remark 2.5.8. We compare the obtained bound (2.28) to Theorem 6 of Bubeck et al. [2013a]. First, the proof reveals that (2.28) holds for all distributions $\underline{\nu}_1 = \bigl(P_0, \mathcal{N}(-\Delta,1)\bigr)$ and $\underline{\nu}_2 = \bigl(P_0, \mathcal{N}(\Delta,1)\bigr)$ where P_0 is a probability distribution with expectation 0. For instance, Bubeck et al. [2013a] considered the Dirac mass δ_0 at 0.

Second, Theorem 6 of Bubeck et al. [2013a] offers the bound

$$\max\bigl\{R_{\psi,\underline{\nu}_1,T},\, R_{\psi,\underline{\nu}_2,T}\bigr\} \geqslant \frac{\log\bigl(T\Delta^2/2\bigr)}{4\Delta}\,. \tag{2.29}$$

Asymptotically, as $T \to +\infty$, our bound (2.28) is smaller by a factor of 2. For small values of T (or small values of Δ), the bound (2.29) is void as the logarithmic term is non-positive, while our bound is always nonnegative. The second argument of the minimum in (2.28) is unimportant, as the regret is always bounded by $T\Delta$.

Proof. We have $R_{\psi,\underline{\nu}_1,T} = \Delta\,\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]$ and $R_{\psi,\underline{\nu}_2,T} = \Delta\,\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]$, so that it suffices to lower bound

$$x = \frac{1}{T}\max\Big\{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big],\ \mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]\Big\}.$$

We assume below that the maximum is given by the first term; otherwise, the proof below should be adapted by exchanging the roles of $\underline{\nu}_1$ and $\underline{\nu}_2$. Inequality (2.6) indicates that

$$\begin{aligned} 2\,T\Delta^2\,x = 2\,\Delta^2\,\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big] &= \mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]\ \mathrm{KL}\big(\mathcal{N}(-\Delta,1),\,\mathcal{N}(\Delta,1)\big) \\ &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}\left(x,\ 1-\frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T}\right). \end{aligned}$$

Given the form of the lower bound in the theorem, which involves a minimum with $T\Delta/2$, we may assume, with no loss of generality, that $x \leqslant 1/2$. Since $\mathrm{kl}(x,\,\cdot\,)$ is increasing on $[x,1]$ and since

$$1-\frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T} \geqslant 1-x \geqslant \frac{1}{2} \geqslant x\,,$$

by definition of x and the assumption $x \leqslant 1/2$, we get

$$2\,T\Delta^2\,x \geqslant \mathrm{kl}(x,\,1-x) = (1-2x)\log\frac{1-x}{x}\,.$$

Note that the case $x = 0$ is excluded by the inequality above. A function study shows that

$$\forall\, x \in (0,1), \qquad (1-2x)\log\frac{1-x}{x} \geqslant \log\frac{1}{2.4\,x}\,.$$

Substituting this lower bound and taking exponents, we are left with studying the inequality

$$\exp\big(2\,T\Delta^2\,x\big) \geqslant \frac{1}{2.4\,x}\,, \qquad \text{or equivalently,} \qquad 2\,T\Delta^2\,x\ \exp\big(2\,T\Delta^2\,x\big) \geqslant \frac{T\Delta^2}{1.2}\,.$$

By definition of the Lambert function W, we rewrite this inequality as $2\,T\Delta^2\,x \geqslant W\big(T\Delta^2/1.2\big)$, which concludes the proof of the first statement.

For the second statement, we note that the property of invariance by translation of the payoffs ensures that

$$x = \frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T} = \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T}\,.$$

Therefore, the fundamental inequality (2.6) directly gives in this case

$$2\,T\Delta^2\,x \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}(x,\,1-x)\,,$$

and we do not need to distinguish whether x is larger than $1/2$ or not. The end of the proof of the first statement of the theorem did not use that $x \leqslant 1/2$ and can still safely be followed for the second statement. □

2.6 A finite-regret algorithm when $\mu^\star$ is known.

In this section, and in this section only, as we are discussing a specific strategy (described below in a box), we will not index the regret, the number of times a given arm is pulled, etc., by the said specific strategy.

We consider the sub-Gaussian framework described in Section 2.1.3 and restrict our attention to the case when $\mu^\star$ is known. We provide a refinement of the results of Bubeck et al. [2013a, Section 3], already known by these authors themselves (see, e.g., [Faure et al., 2015]). The algorithm considered below is inspired by Algorithm 1 of Bubeck et al. [2013a]. For each $t \geqslant 1$ and $a \in \{1, \dots, K\}$ such that $N_a(t) \geqslant 1$, we denote by

$$\widehat{\mu}_{a,t} = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s = a\}}$$

the empirical mean of the rewards obtained between rounds 1 and t when playing arm a.

Algorithm 7: An algorithm with bounded regret, thanks to the knowledge of $\mu^\star$

Bandit problem: $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ where each ν_a is sub-Gaussian in the sense of (2.4)

Parameters: the value of $\mu^\star = \max_{a=1,\dots,K} \mu_a$

For: each $t \in \{1, \dots, K\}$, **do:** play arm t.

For: each round $t \geqslant K+1$,

1. Let $\mathcal{C}_t = \left\{ a \in \{1, \dots, K\} \ : \ \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\dfrac{4 \log N_a(t-1)}{N_a(t-1)}} \right\}$ be the set of candidate arms;
2. If $\mathcal{C}_t \neq \emptyset$, play an arm A_t at random in $\mathcal{C}_t$, update $t := t+1$;
3. If $\mathcal{C}_t = \emptyset$, play $A_t = 1$, $A_{t+1} = 2$, $\dots$, $A_{t+K} = t+K-1$, update $t := t+K$.

We use the notation introduced before (2.1), but, as indicated above, without the indexations in the considered strategy.

Theorem 2.6.1. *For all bandit problems $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ where each distribution ν_a is sub-Gaussian in the sense of (2.4), the regret of the algorithm above is bounded by*

$$R_{\underline{\nu},T} \leqslant \sum_{a:\Delta_a>0} \left(\frac{36 \log(17/\Delta_a)}{\Delta_a} + 3\Delta_a \right).$$

Proof. We fix an optimal arm $a^\star$. In view of (2.1), it suffices to bound $\mathbb{E}_\nu\big[N_a(T)\big]$ for each suboptimal arm a. Each arm is played once between 1 and K. For all $t \geqslant K+1$, a suboptimal arm a can only be played if $a \in \mathcal{C}_t$ (step 2 of the second for loop) or if we are in a sequence where each arm is played successfully (step 3 of the second for loop). In the latter case, the set of candidate arms at round $t-a+1$ was empty. It did not contain $a^\star$. This optimal arm is played also once in the sequence of pulls corresponding to step 3, at time $t-a+a^\star+1$. At time $t-a+a^\star$ we still had $N_{a^\star}(t-a+a^\star) = N_{a^\star}(t-a+1)$, so that the condition for being a candidate was violated as well:

$$\widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\log N_a(t-a+a^\star)}{N_a(t-a+a^\star)}}\,.$$

All in all, we proved the inclusion: for $t \geqslant K+1$,

$$\begin{aligned}\{A_t = a\} \subseteq \quad & \left\{ A_t = a \;\text{ and }\; \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\frac{4\log N_a(t-1)}{N_a(t-1)}} \right\} \\ & \cup \left\{ A_{t-a+a^\star} = a^\star \;\text{ and }\; \widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\log N_a(t-a+a^\star)}{N_a(t-a+a^\star)}} \right\}.\end{aligned}$$

We now only sketch the next argument, as we proceed similarly to all multi-armed bandit analyses, by resorting to Doob's optional sampling theorem, which asserts that the rewards Y_s obtained at those rounds s when $A_s = a$ are independent and identically distributed according to ν_a. We denote by $\overline{\mu}_{a,n}$ the empirical average of the first n rewards obtained by arm a during the game. Then,

$$\begin{aligned}\mathbb{E}_\nu\big[N_a(T)\big] &\leqslant 1 + \sum_{t=K+1}^{T} \mathbb{P}\left\{ A_t = a \;\text{ and }\; \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\frac{4\log N_a(t-1)}{N_a(t-1)}} \right\} \\ &\qquad + \sum_{t=K+1}^{T} \mathbb{P}\left\{ A_{t-a+a^\star} = a^\star \;\text{ and }\; \widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\log N_a(t-a+a^\star)}{N_a(t-a+a^\star)}} \right\} \\ &\leqslant 1 + \sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu^\star > -\sqrt{\frac{4\log n}{n}} \right\} + \sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a^\star,n} - \mu^\star \leqslant -\sqrt{\frac{4\log n}{n}} \right\}.\end{aligned} \tag{2.30}$$

As indicated already in Bubeck et al. [2013a], for each arm a, the sub-Gaussian assumption on ν_a, together with a Crámer–Chernoff bound, indicates that for all $n \geqslant 1$ and all $\varepsilon > 0$,

$$\max\Big\{ \mathbb{P}\big\{\overline{\mu}_{a,n} - \mu_a \geqslant \varepsilon\big\},\; \mathbb{P}\big\{\overline{\mu}_{a,n} - \mu_a \leqslant -\varepsilon\big\} \Big\} \leqslant \exp\left(-\frac{n\varepsilon^2}{2}\right). \tag{2.31}$$

We substitute this inequality in the bound (2.30) obtained above. On the one hand, for $a^\star$,

$$\sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a^\star,n} - \mu^\star \leqslant -\sqrt{\frac{4\log n}{n}} \right\} \leqslant \sum_{n\geqslant 1} n^{-2} \leqslant 2\,. \tag{2.32}$$

On the other hand, for a, we rewrite $\mu^\star = \mu_a + \Delta_a$ and get

$$\sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu^\star > -\sqrt{\frac{4\log n}{n}} \right\} = \sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu_a > \Delta_a - \sqrt{\frac{4\log n}{n}} \right\}.$$

To upper bound the latter sum, we denote by n_0 the smallest integer $k \geqslant 3$, if it exists, such that:

$$\Delta_a - \sqrt{\frac{4\log k}{k}} \geqslant \frac{\Delta_a}{2}, \qquad \text{that is,} \qquad \sqrt{\frac{4\log k}{k}} \leqslant \frac{\Delta_a}{2}. \tag{2.33}$$

As $x \mapsto \sqrt{(\log x)/x}$ is decreasing on $[3, +\infty)$, we have

$$\forall n \geqslant n_0, \qquad \Delta_a - \sqrt{\frac{4\log n}{n}} \geqslant \frac{\Delta_a}{2},$$

and thus

$$\sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu_a > \Delta_a - \sqrt{\frac{4\log n}{n}} \right\} \leqslant n_0 - 1 + \sum_{n\geqslant n_0} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu_a > \frac{\Delta_a}{2} \right\}.$$

Note that the above inequality also holds with $n_0 = 2$ when no $k \geqslant 3$ satisfies (2.33). We use (2.31) and a comparison to an integral to get

$$\sum_{n\geqslant n_0} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu_a > \frac{\Delta_a}{2} \right\} \leqslant \sum_{n\geqslant n_0} \exp\left(-\frac{n\Delta_a^2}{8}\right) \leqslant \int_{n_0-1}^{+\infty} \exp\left(-\frac{x\Delta_a^2}{8}\right) \mathrm{d}x \leqslant \frac{8}{\Delta_a^2}.$$

Substituting the above bounds and (2.32) into (2.30), we showed so far that

$$\mathbb{E}_\nu\big[N_a(T)\big] \leqslant n_0 + 2 + \frac{8}{\Delta_a^2}.$$

The proof is concluded by upper bounding n_0, based on (2.33). If $\Delta_a \leqslant 4\sqrt{(\log 3)/3}$, then the n_0 defined in (2.33) exists. In this case, we denote by $x_0 \in [3, +\infty)$ the real number such that

$$\sqrt{\frac{4\log x_0}{x_0}} \leqslant \frac{\Delta_a}{2} \qquad \text{that is,} \qquad x_0 = \frac{16\log x_0}{\Delta_a^2}.$$

We have $n_0 = \lceil x_0 \rceil \leqslant x_0 + 1$. Since

$$x_0 = \frac{16\log x_0}{\Delta_a^2} = \frac{32\log(4/\Delta)}{\Delta_a^2} + \frac{16}{\Delta_a^2}\log\big(\log x_0\big),$$

we suspect that x_0 should not be too much larger than $32\log(4/\Delta)/\Delta_a^2$. Indeed, using the inequality $\log(u) \leqslant u$, we see that

$$x_0 = \frac{16\log x_0}{\Delta_a^2} = \frac{160\log x_0^{1/10}}{\Delta_a^2} \leqslant \frac{160\, x_0^{1/10}}{\Delta_a^2}, \qquad \text{thus} \qquad x_0 \leqslant \left(\frac{160}{\Delta_a^2}\right)^{10/9}.$$

Therefore,

$$x_0 = \frac{16 \log x_0}{\Delta_a^2} \leqslant \frac{16}{\Delta_a^2} \log\left(\frac{160}{\Delta_a^2}\right)^{10/9} \leqslant \frac{16 \times (10/9) \times 2}{\Delta_a^2} \log \frac{13}{\Delta^2} \leqslant \frac{36}{\Delta_a^2} \log \frac{13}{\Delta^2}\,.$$

When the n_0 defined in (2.33) does not exist and we take $n_0 = 2$, we may still bound n_0 by 1 plus the bound above on x_0 (as the latter is larger than 1). The theorem follows, after substitution of all the bounds, together with the inequality $8 \leqslant 36 \log(17) - 36 \log(13)$. □

Chapter 3

kl-UCB Algorithms for Exponential Families

In collaboration with Aurélien Garivier.

Contents

3.1 Introduction

For regret minimization in stochastic bandit problems, two notions of time-optimality coexist. On the one hand, one may consider a fixed model: the famous lower bound by Lai and Robbins [1985] showed that the regret of any consistent strategy should grow at least as $C(\mu)\log(T)\big(1-o(1)\big)$ when the horizon T goes to infinity. Here, $C(\mu)$ is a constant depending solely on the model. A strategy with a regret upper-bounded by $C(\mu)\log(T)\big(1+o(1)\big)$ will be called in this chapter *asymptotically optimal*. Lai and Robbins provided a first example of such a strategy in their seminal work. Later, Garivier and Cappé [2011] and Maillard et al. [2011] provided finite-time analysis for variants of the UCB algorithm (see Agrawal [1995], Burnetas and Katehakis [1996], Auer et al. [2002a]) which imply asymptotic optimality. Since then, other algorithms like Bayes-UCB [Kaufmann et al., 2012] and Thompson Sampling [Korda et al., 2013] have also joined the family.

On the other hand, for a fixed horizon T one may assess the quality of a strategy by the greatest regret suffered in all possible bandit models. If the regret of a bandit strategy is upper-bounded by $C'\sqrt{KT}$ (the optimal rate: see Auer et al. [2002b] and Cesa-Bianchi and Lugosi [2006]) for some numeric constant C', this strategy is called *minimax optimal*. The PolyINF and the MOSS strategies by Audibert and Bubeck [2009] were the first proved to be minimax optimal for bandit problem with bounded rewards (cf. Chapter 4).

Hitherto, as far as we know, no algorithm was proved to be *at the same time* asymptotically and minimax optimal. Two limited exceptions may be mentioned: the case of two Gaussian arms is treated by Garivier et al. [2016]; and the OC-UCB algorithm of Lattimore [2015] is proved to be minimax-optimal and almost problem-dependent optimal for Gaussian multi-armed bandit problems. Notably, the OC-UCB algorithm satisfies another worthwhile property of *finite-time instance near-optimality*, see Section 2 of Lattimore [2015] for a detailed discussion. In the same line of works we can cite the AdaUCB from Lattimore [2018] which is minimax optimal, problem-dependent optimal and finite-time instance near-optimal for Gaussian multi-armed bandit problems. Nevertheless the analysis heavily relies on some particular properties of the Gaussian distributions and it is not clear how to adapt this analysis to Bernoulli distributions for example.

Contributions. In this chapter, we put forward the kl-UCB^{++} algorithm, a slightly modified version of kl-UCB^{+} algorithm discussed in Garivier and Cappé [2011] as an empirical improvement of UCB, and analyzed in Kaufmann [2016]. This bandit strategy is designed for some exponential distribution families, including for example Bernoulli and Gaussian laws. It borrows from the MOSS algorithm of Audibert and Bubeck [2009] the idea to divide the horizon by the number of arms in order to reach minimax optimality. We prove that it is at the same time asymptotically and minimax optimal. This work thus merges the progress which has been made in different directions towards the understanding of the optimism principle, finally reconciling the two notions of time-

optimality.

Insofar, our contribution answers a very simple and natural question. The need for simultaneous minimax- and problem-dependent optimality could only be addressed in very limited settings by means that could not be generalized to the framework adopted in this chapter. Indeed, for a given horizon T, the worst problem depends on T: it involves arms separated by a gap of order $\sqrt{K/T}$. Treating the T-dependent problems correctly for all T appears as a quite different task than catching the optimal, problem-dependent speed of convergence for every fixed bandit model. We show in this chapter that the two goals can indeed be achieved simultaneously.

Combining the two notions of optimality requires a modified exploration rate. We stick as much as possible to existing algorithms and methods, introducing just what is necessary to obtain the desired results. Starting from that of kl-UCB (so as to have a tight asymptotic analysis), one has to completely cancel the exploration bonus of the arms that have been drawn roughly T/K times. The consequence is very slight and harmless in the case where the best arm is much better than the others, but essential in order to minimize the regret in the worst case where the best arm is barely distinguishable from the others.

We present a general yet simple proof, combining the best elements of the above-cited sources which are simplified as much as possible and presented in a unified way. To this end, we develop new deviation inequalities, improving the analysis of the different terms contributing to the regret. This analysis is made in the framework which we believe is the best compromise between simplicity and generality (simple exponential families). This permits us to treat, among others, the Bernoulli and the Gaussian case at the same time. More fundamentally, this appears to us as the right, simple framework for the analysis, which emphasizes what is really required to have simple lower- and upper-bounds (the possibility to make adequate changes of measure, and Chernoff-type deviation bounds).

3.2 One Parameter Exponential Families

In this chapter, each arm is assumed to be a probability distribution of some canonical one-dimensional exponential family ν_θ indexed by $\theta \in \Theta$. The probability law ν_θ is assumed to be absolutely continuous with respect to a dominating measure ρ on $\mathbb{R}$, with a density given by

$$\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\rho}(x) = \exp(x\theta - b(\theta)), \quad \text{where } b(\theta) = \log \int_{\mathbb{R}} e^{x\theta} \mathrm{d}\rho(x) \text{ and } \Theta = \left\{\theta \in \mathbb{R} : \ b(\theta) < +\infty\right\}.$$

It is well-known that b is convex, twice differentiable on Θ, that $b'(\theta) = E(\nu_\theta)$ and $b''(\theta) = V(\nu_\theta) > 0$, respectively the mean and the variance of the distribution ν_θ. The family can thus be parametrized by the mean $\mu = b'(\theta)$, for $\mu \in I := b'(\Theta)$. The Kullback-Leibler divergence between two distributions is $\mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - b'(\theta)(\theta' - \theta)$. This permits to define the following divergence on the set of arm expectations: if $\mu = E(\nu_\theta)$ and $\mu' = E(\nu_{\theta'})$ then

$$d(\mu, \mu') := \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b^*(\mu) - b^*(\mu') - b^{*\prime}(\mu')(\mu - \mu'),$$

where b^* is the Fenchel conjugate of b. For a minimax analysis, we need to restrict the set of means to a bounded interval: we suppose that each arm ν_θ satisfies $\mu = b'(\theta) \in [\mu^-, \mu^+] \subset I$ for two fixed real numbers μ^+, μ^-. Our analysis requires some kind of Pinsker's inequality; we therefore assume that the variance is bounded in the exponential family: there exists $V > 0$ such that

$$\sup_{\mu \in I} b''\big(b'^{-1}(\mu)\big) = \sup_{\mu \in I} V\big(\nu_{b'^{-1}(\mu)}\big) \leqslant V < +\infty \,.$$

This implies that for all $\mu, \mu' \in I$,

$$d(\mu, \mu') \geqslant \frac{1}{2V}(\mu - \mu')^2 \,. \tag{3.1}$$

In the sequel, we denote by $\mathcal{F}$ the set of bandit problems $\underline{\nu}$ satisfying these assumptions. This setting includes in particular the following example

Example 3.2.1. (Bernoulli distribution). $\Theta = \mathbb{R}$, $b(\theta) = \log\big(1 + \exp(\theta)\big)$, $I = (0, 1)$, $V = 1/4$,

$$d(\mu, \mu') = \mathrm{kl}(\mu, \mu').$$

Example 3.2.2. (Gaussian distribution with known variance σ^2). $\Theta = \mathbb{R}$, $b(\theta) = \sigma^2\theta/2$, $I = \mathbb{R}$, $V = \sigma^2$,

$$d(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2).$$

3.3 Two criteria of optimality

In this section we describe two criteria of optimality: asymptotic and minimax optimality, then we present an algorithm kl-UCB^{++} which reaches these two notions of optimalty.

3.3.1 Lower Bounds on the Regret

We recall the asymptotic lower bound of Burnetas and Katehakis [1996] and define an asymptotically optimal strategy. This lower bound is a particular case of Theorem 2.2.5.

Theorem 3.3.1. *(Asymptotic lower bound) For all uniformly fast convergent strategies (see Definition 2.2.4), for all bandit problems $\underline{\nu} \in \mathcal{F}$, for all suboptimal arms a,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log T} \geqslant \frac{1}{d(\mu_a, \mu^\star)} \,.$$

Definition 3.3.2. A strategy is *asymptotically optimal* on $\mathcal{F}$, if for all bandit problems $\underline{\nu} \in \mathcal{F}$, for all sub-optimal arms a,

$$\limsup_{T \to \infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\log(T)} \leqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)} \,.$$

In the same spirit of the minimax (or distribution-free) lower bound of Auer et al. [2002b] on can prove a similar lower bound (see Section 2.5.2 for a proof) for the family of bandit problems $\mathcal{F}$. Let $V' := \sup_{\mu\in[\mu^-,\mu^+]} b''\big(b'^{-1}(\mu)\big)$ be the maximum of the variance in the one parameter exponential family over the interval $[\mu^-, \mu^+]$.

Theorem 3.3.3. *(Minimax lower bound). There exists a constant ε_0 that depends uniquely on the family of bandit problems $\mathcal{F}$ and the two endpoints of the interval $[\mu^-, \mu^+]$ such that for all strategies,*

$$\sup_{\underline{\nu}\in\mathcal{F}} R_{T,\underline{\nu}} \geqslant \frac{1}{16}\min\Big\{\sqrt{V'KT},\, \varepsilon_0^2 T\Big\}\,.$$

One could not expect to have the constant V instead of V' in the lower bound since we restrain our-self to arms with parameter μ lying within the interval $[\mu^-, \mu^+]$.

Definition 3.3.4. A strategy is minimax optimal on $\mathcal{F}$ if there exists a constant C such that for all bandit problem $\underline{\nu} \in \mathcal{F}$, for all T,

$$R_{T,\underline{\nu}} \leqslant C\sqrt{KT}\,.$$

Note that the notion of minimax optimality is defined here up to a multiplicative constant, in contrast to the definition of (problem-dependent) asymptotic optimality.

3.3.2 An Asymptotically and Minimax Optimal Algorithm

We denote by $\widehat{\mu}_{a,n}$ the empirical mean of the first n rewards from arm a, after t rounds it is

$$\widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)} = \frac{1}{N_a(t)}\sum_{s=1}^{t} Y_s\,\mathbb{1}_{\{A_s=a\}}\,.$$

Algorithm 8: Generic kl-UCB algorithm.

Parameters: A function $f : \mathbb{N}\times\mathbb{N}\times\mathbb{N} \to \mathbb{R}^+$

Initialization: Pull each arm of $\{1, .., K\}$ once.

For $t = K$ to $T-1$, **do**

1. Compute for each arm a the quantity

$$U_a^{\mathrm{kl}}(t) = \sup\left\{\mu \in [0,1] \;:\; d\big(\widehat{\mu}_a(t), \mu\big) \leqslant \frac{f\big(N_a(t), t, T\big)}{N_a(t)}\right\}. \tag{3.2}$$

2. Play $A_t \in \arg\max_{a\in\{1,..,K\}} U_a^{\mathrm{kl}}(t)$.

The kl-UCB^{++} algorithm is a slight modification of algorithm kl-UCB^{+} of Garivier and Cappé [2011] analyzed by Kaufmann [2016] which uses the exploration function

$$f(n,t,T) = \log\left(\frac{T}{n}\right). \tag{3.3}$$

It uses the exploration function g with an extra factor K given by

$$f(n,t,T) = g(n) := \log_+\left(\frac{T}{Kn}\right), \tag{3.4}$$

where $\log_+(x) := \max\big(\log(x), 0\big)$. The exploration function is the same as the one of MOSS algorithm and therefore kl-UCB^{++} reduce to MOSS algorithm when the divergence is the euclidean squared distance. In fact, it borrows from kl-UCB algorithm of Cappé et al. [2013] the divergence and from MOSS algorithm the exploration function. The following results state that the kl-UCB^{++} algorithm is *simultaneously* minimax and asymptotically optimal.

Theorem 3.3.5 (Minimax optimality)**.** *For any bandit model $\underline{\nu} \in \mathcal{F}$, the expected regret of the* kl-*UCB^{++} algorithm is upper-bounded as*

$$R_T \leqslant 33\sqrt{VKT} + (\mu^+ - \mu^-)K\,. \tag{3.5}$$

Theorem 3.3.6 (Asymptotic optimality)**.** *For any bandit model $\underline{\nu} \in \mathcal{F}$, for any suboptimal arm a and any δ such that $22VK/T \leqslant \delta^2 \leqslant (\mu^\star - \mu_a)^2/9$,*

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{\mathrm{kl}(\mu_a + \delta, \mu^\star - \delta)} + O_\delta\big(\operatorname{loglog}(T)\big) \tag{3.6}$$

(see the end of the proof in Section 3.3.4 for an explicit bound). In particular,

$$\limsup_{T\to\infty} \frac{\mathbb{E}_a\big[N_a(T)\big]}{\log(T)} \leqslant \frac{1}{\mathrm{kl}(\mu_a, \mu^\star)}\,. \tag{3.7}$$

Note that in (3.5), it is a new instance of the algorithm for each T since we need to know the horizon T. The same remark holds for the bound of Equation 3.6. For an anytime minimax analysis of the MOSS algorithm (corresponding to the choice of Gaussian arms with variance $\sigma^2 = 1/4$) see the proof of Proposition 4.5.3. But the algorithm does not need to know in advance the endpoints of the interval $[\mu^-, \mu^+]$, that's why we get the constant V and not V' inside the bound of Equation 1.8. It could be possible to obtain the right constant V' by injecting the knowlegde of μ^- and μ^+ in the construction on the index $U_a^{\mathrm{kl}}(t)$.

Theorems 3.3.5 and 3.3.6 are proved in Sections 3.3.3 and 3.3.4 respectively. The main differences between the two proofs are discussed at the beginning of Section 3.3.3. Note that the two regret bounds of Theorems 3.3.5 and 3.3.6 also apply to all $[0,1]$-valued bandit models, with the value $V = 1/4$, as the deviations of $[0,1]$-valued random

variables are dominated by those of a Bernoulli distribution with the same mean (this is discussed for example by Cappé et al. [2013]). However, the kl-UCB^{++} algorithm is not asymptotically optimal then: the regret bound in $\log(T)/\mathrm{kl}(\mu_a, \mu^*)$ is not optimal in that case. See Chapter 4 for an asymptotic and minimax optimal algorithm for $[0,1]$-valued bandit models based on the empirical-likelihood method.

3.3.3 Proof of Theorem 3.3.5

This proof merges merges ideas presented by Bubeck and Liu [2013] for the analysis of the MOSS algorithm and from the analysis of kl-UCB by Cappé et al. [2013] (see also Kaufmann [2016]). It is divided into the following steps:

Decomposition of the regret. Let $a^\star$ be the index of an optimal arm. Since by definition of the strategy $U_{a^\star}(t) \leqslant U_{A_{t+1}}(t)$ for all $t \geqslant K-1$, the regret can be decomposed as follows:

$$R_T \leqslant K(\mu^+ - \mu^-) + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big]}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{E}\big[U_{A_{t+1}}(t) - \mu_{A_{t+1}}\big]}_{B} . \tag{3.8}$$

For the first term A, as in the proof of MOSS algorithm, we carefully upper bound the probability that appears inside the integral thanks to a 'peeling trick'. The second term B is easier to handle since we can reduce the index to UCB-like-index thanks to the Pinsker inequality (3.1) and proceed as Bubeck and Liu [2013].

Step 1: Upper-bounding A. Term A is concerned with the optimal arm $a^\star$ only. Two words of intuition: since $U_{a^\star}(t)$ is meant to be an upper confidence bound for $\mu^\star$, this term should not be too large, at least as long as the the confidence level controlled by function g is large enough – but when the confidence level is low, the number of draws is large and deviations are unlikely.

Therefore to upper-bound term A we separate two cases depending on whether $N_a^\star(t)$ is greater or lower than $N := T/K$

$$\mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big] \leqslant \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\left\{N_a^\star(t)<N\right\}}\Big] + \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\left\{N_a^\star(t)\geqslant N\right\}}\Big] . \tag{3.9}$$

For the second term, since $N_{a^\star}(t)$ is large enough, we just need to use the deviations of the mean. Using the maximal inequality, recalled in (3.49), one obtains

$$\begin{aligned}\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\{N_a^\star(t)\geqslant N\}}\Big] &\leqslant \mathbb{E}\big[\max_{n\geqslant N}(\mu^\star - \widehat{\mu}_{a^\star,n})\big] \\ &\leqslant \sqrt{\frac{\pi}{2}}\sqrt{\frac{V}{N}} = \sqrt{\frac{\pi}{2}}\sqrt{\frac{KV}{T}}\end{aligned} \tag{3.10}$$

The first term can be upper-bounded thanks to a 'peeling trick' as in the proof of MOSS. We use the grid $N/\beta^{l+1} \leqslant N_a^\star \leqslant N/\beta^l$, where the real $\beta > 1$ will be chosen later. We get using the peeling trick, then integrating the deviations

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\left\{N_a^\star(t)<N\right\}}\Big] \leqslant \sum_{l=0}^{+\infty} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)^+ \mathbb{1}_{\left\{N/\beta^{l+1}\leqslant N_a^\star(t)\leqslant N/\beta^l\right\}}\Big] \tag{3.11}$$

$$\leqslant \sum_{l=0}^{+\infty} \int_{u=0}^{+\infty} \underbrace{\mathbb{P}\bigg(\exists \frac{N}{\beta^{l+1}} \leqslant n \leqslant \frac{N}{\beta^l},\, \mu^\star - U_{a^*,n} \geqslant u\bigg)}_{:=A^l} \mathrm{d}u \tag{3.12}$$

Thanks to the definition of the index we can rewrite the probability appearing in the integral. On the event $\{U_{a^\star,n} \leqslant \mu^\star - u\}$, we have that $\widehat{\mu}_{a^\star,n} \leqslant U_{a^\star,n} \leqslant \mu^\star - u < \mu^\star$. Consequently, it holds that, defining $d_+(p,q) := d(p,q)\mathbb{1}_{\{p\leqslant q\}}$,

$$\begin{aligned} A^l &\leqslant \mathbb{P}\bigg(\exists \frac{N}{\beta^{l+1}} \leqslant n \leqslant \frac{N}{\beta^l},\, n d_+(\widehat{\mu}_{a^*,n}, \mu^\star - u) \geqslant g(n)\bigg) \\ &\leqslant \mathbb{P}\bigg(\exists \frac{N}{\beta^{l+1}} \leqslant n \leqslant \frac{N}{\beta^l},\, d_+(\widehat{\mu}_{a^*,n}, \mu^\star) \geqslant \frac{g(n)}{n} + \frac{u^2}{2V}\bigg)\,. \end{aligned} \tag{3.13}$$

Using again a maximal inequality, recalled in Lemma 3.5.5, but this time for the deviations of $d_+(\widehat{\mu}_{a^*,n}, \mu^\star)$, we have

$$A_l \leqslant \exp\bigg(-\frac{1}{\beta} g(N/\beta^l) - \frac{N}{\beta^{l+1}}\frac{u^2}{2V}\bigg)\,.$$

Injecting this upper-bound in Inequality (3.11) then replacing g and $N = T/K$ by their values, leads to

$$\begin{aligned} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\{N_a^\star<N\}}\Big] &\leqslant \sum_{l=0}^{+\infty} e^{g(N\beta^l)/\beta} \int_0^{+\infty} e^{-Nu^2/(2V\beta^{l+1})}\mathrm{d}u \\ &= \sqrt{\frac{KV}{T}}\sqrt{\frac{\pi}{2}} \sum_{l=0}^{+\infty} \sqrt{\beta} e^{-l\log(\beta)(1/\beta - 1/2)}\,. \end{aligned}$$

Choosing $\beta = 3/2$ in order to make converge the sum, gives

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}(t)\big)\mathbb{1}_{\{N_a^\star<N\}}\Big] \leqslant \sqrt{\frac{KV}{T}}\sqrt{\frac{\pi}{2}} 19\,. \tag{3.14}$$

Summing over t from K to $T-1$ (3.14) and (3.10) yields the bound on term A

$$A \leqslant 20\sqrt{\frac{\pi}{2}}\sqrt{VKT}\,. \tag{3.15}$$

Step 2: Upper-bounding B. Term B is of different nature, since typically $U_{A_{t+1}}(t) > \mu_{A_{t+1}}$. We define $\delta = \sqrt{VK/T}$; since the bound (3.5) is otherwise trivial, we assume in the sequel that $\delta \leqslant 1$. However, as for the term A, we first reduce the problem to the upper-bounding of a probability:

$$B \leqslant T\delta + \sum_{t=K}^{T-1} \mathbb{E}\big[(U_{A_{t+1}}(t) - \mu_{A_{t+1}} - \delta)^+\big] \,. \tag{3.16}$$

To get rid of the randomness of $N_{A_{t+1}}(t)$ we use the pessimistic trajectorial upper bound from Bubeck and Liu [2013]

$$\sum_{t=K}^{T-1} (U_{A_{t+1}}(t) - \mu_{A_{t+1}} - \delta)^+ \leqslant \sum_{n=1}^{T} \sum_{a=1}^{K} (U_{a,n} - \mu_a - \delta)^+ \,.$$

In addition, we simplify the upper-bound thanks to our assumption (3.1) that some Pinsker type inequality is available:

$$U_{a,n} \leqslant B_{a,n} := \widehat{\mu}_{a,n} + \sqrt{2V \frac{g(n)}{n}} \,. \tag{3.17}$$

Hence, B can be upper-bounded as

$$B \leqslant T\delta + \sum_{a=1}^{K} \sum_{n=1}^{T} \mathbb{E}\big[(B_{a,n} - \mu_a - \delta)^+\big] \,. \tag{3.18}$$

Then, we need only to upper bound $\sum_{n=1}^{T} \mathbb{E}\big[(B_{a,n} - \mu_a - \delta)^+\big]$ for each arm $a \in \{1, \dots, K\}$. We cut the sum at the critical sample size $N = T/K$ when the exploration bonus of $B_{a,n}$ vanishes. Thus, if $n < N$, we have

$$(B_{a,n} - \mu_a - \delta)^+ \leqslant (\widehat{\mu}_{a,n} - \mu_a - \delta)^+ + \sqrt{\frac{2Vg(n)}{n}} \,,$$

else $n \geqslant N$, we get

$$(B_{a,n} - \mu_a - \delta)^+ \leqslant (\widehat{\mu}_{a,n} - \mu_a - \delta)^+ \,.$$

Combining this two inequalities leads to

$$\sum_{n=1}^{T} \mathbb{E}\big[(B_{a,n} - \mu_a - \delta)^+\big] \leqslant \sum_{n=1}^{T} \mathbb{E}\big[(\widehat{\mu}_{a,n} - \mu_a - \delta)^+\big] + \sum_{1 \leqslant n \leqslant N} \sqrt{\frac{2Vg(n)}{n}} \,. \tag{3.19}$$

For the first sum we use Inequality (3.50) and get

$$\begin{aligned}
\sum_{n=1}^{T} \mathbb{E}\big[(\widehat{\mu}_{a,n} - \mu_a - \delta)^+\big] &\leqslant \sum_{n=1}^{T} \sqrt{\frac{V\pi}{2n}} e^{-n\delta^2/(2V)} \\
&\leqslant \int_0^{+\infty} \sqrt{\frac{V\pi}{2x}} e^{-x\delta^2/(2V)} \mathrm{d}x \\
&= \frac{V}{\delta} \sqrt{\pi} \int_0^{+\infty} x^{-1/2} e^{-x} \mathrm{d}x = \frac{V}{\delta} \pi \,.
\end{aligned}$$

The second sum is simpler to handle, replacing g by its value, one obtains

$$\begin{aligned}\sum_{1\leqslant n\leqslant N}\sqrt{\frac{2Vg(n)}{n}} &\leqslant \sqrt{2V}\int_0^{+\infty}\sqrt{\frac{\log\left(T/(Kx)\right)}{x}}\mathrm{d}x\\ &= \sqrt{\frac{2VT}{K}}\int_0^{+\infty}\sqrt{\log(x)}x^{-3/2}\mathrm{d}x\\ &= \sqrt{\frac{4\pi VT}{K}}\,.\end{aligned}$$

Thus using this two upper-bounds in (3.19) and leads to

$$B \leqslant T\delta + \frac{V}{\delta}\pi + \sqrt{4\pi VTK}$$

Replacing $\delta = \sqrt{VK/T}$ by its value allows us to conclude for term B

$$B \leqslant (1+\pi+\sqrt{4\pi})\sqrt{VKT}\,. \tag{3.20}$$

Conclusion of the proof. It just remains to plug Inequalities (3.15) and (3.20) into Equation (3.8):

$$\begin{aligned}A+B &\leqslant \left(20\sqrt{\frac{\pi}{2}}+1+\pi+\sqrt{4\pi}\right)\sqrt{VKT}\\ &\leqslant 33\sqrt{VKT}\,,\end{aligned}$$

which concludes the proof.

3.3.4 Proof of Theorem 3.3.6

The analysis of asymptotic optimality shares many elements with the minimax analysis, with some differences however. The decomposition of the regret into two terms A and B is similar, but localized on a fixed sub-optimal arm $a \in \{1,\dots,K\}$: we analyze the number of draws of a and not directly the regret (and we do not need to integrate the deviations at the end). We proceed roughly as in the proof of Theorem 3.3.5 for term A, which involves the deviations of an optimal arm. For term B, which stands for the behavior of the sub-optimal arm a, a different (but classical) argument is used, as one cannot simply use the Pinsker-like Inequality (3.1) if one wants to obtain the correct constant (and thus asymptotic optimality).

Decomposition of $\mathbb{E}\big[N_a(T)\big]$. If arm a is pulled at time $t+1$, then by definition of the strategy $U_{a^*}(t) \leqslant U_a(t)$ for any index a^* of an optimal arm. Thus,

$$\begin{aligned}\big\{A_{t+1}=a\big\} &\subseteq \big\{\mu^\star - \delta \geqslant U_a(t)\big\} \cup \big\{\mu^\star-\delta < U_a(t) \text{ and } A_{t+1}=a\big\}\\ &\subseteq \big\{\mu^\star - \delta \geqslant U_{a^*}(t)\big\} \cup \big\{\mu^\star-\delta < U_a(t) \text{ and } A_{t+1}=a\big\}\,.\end{aligned}$$

As a consequence,

$$\mathbb{E}\big[N_a(T)\big] \leqslant 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(U_{a^*}(t) \leqslant \mu^\star - \delta\big)}_{\text{A}} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(\mu^\star - \delta < U_a(t) \text{ and } A_{t+1} = a\big)}_{\text{B}}\,, \quad (3.21)$$

and it remains to bound each of these terms.

Step 1: Upper-bounding term A. As in the proof of Theorem 3.3.5, with $N := T/K$, we write

$$\mathbb{P}\big(U_{a^*}(t) \leqslant \mu^\star - \delta\big) \leqslant \underbrace{\mathbb{P}\big(\exists 1 \leqslant n < N, \quad d_+(\widehat{\mu}_{a^\star,n}, \mu^\star - \delta) \geqslant g(n)/n\big)}_{A_1} + \underbrace{\mathbb{P}\big(\exists N \leqslant n \leqslant T, \quad \widehat{\mu}_{a^\star,n} \leqslant \mu^\star - \delta\big)}_{A_2}\,. \quad (3.22)$$

Using Lemma 3.5.3 we can rewrite the term A_1 with δ outside the Kullback-Leibler divergence

$$A_1 \leqslant \mathbb{P}\big(\exists 1 \leqslant n < N, \quad d_+(\widehat{\mu}_{a^\star,n}, \mu^\star) \geqslant g(n)/n + \delta^2/(2V)\big)\,.$$

Then the union bound and Inequality (3.46) lead to the following (rather crude) upper bound of term A_1

$$\begin{aligned} A_1 &\leqslant \sum_{1 \leqslant n \leqslant N} \mathbb{P}\big(d_+(\widehat{\mu}_{a^\star,n}, \mu^\star) \geqslant g(n)/n + \delta^2/(2V)\big) \\ &\leqslant \frac{T}{K} \sum_{n=1}^{\infty} n e^{-n\delta^2/(2V)} \leqslant \frac{T}{K} \frac{(2V)^2}{\delta^4}\,. \end{aligned} \quad (3.23)$$

Thanks to the maximal inequality recalled in Inequality (3.47), it holds that

$$A_2 \leqslant e^{-\delta^2 N/(2V)} = e^{-\delta^2 T/(2KV)}\,. \quad (3.24)$$

Putting Equations (3.22) to (3.24) together yields:

$$A \leqslant \frac{(2V)^2}{K\delta^4} + T e^{-\delta^2 T/(2KV)}\,. \quad (3.25)$$

Step 2: Upper-bounding B. Thanks to the definition of $U_a(t)$ it holds that

$$\big\{\mu^\star - \delta < U_a(t) \text{ and } A_{t+1} = a\big\} \subseteq \Big\{d\big(\widehat{\mu}_a(t), \mu^\star - \delta\big) \leqslant g\big(N_a(t)\big)/N_a(t) \text{ and } A_{t+1} = a\Big\}$$

Together with the following classical bandit reasoning, this yields:

$$\begin{aligned} B &\leqslant \sum_{t=K}^{T-1} \mathbb{P}\big(d\big(\widehat{\mu}_a(t), \mu^\star - \delta\big) \leqslant g\big(N_a(t)\big)/N_a(t) \text{ and } A_{t+1} = a\big) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}\big(d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant g(n)/n\big) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}\Big(d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant \log(T/K)/n\Big)\,, \end{aligned} \tag{3.26}$$

as function g is non-increasing. Now, let $n(\delta)$ be the integer defined as

$$n(\delta) = \left\lceil \frac{\log(T/K)}{d(\mu_a + \delta, \mu^\star - \delta)} \right\rceil .$$

Recall that by assumption $\delta < (\mu^\star - \mu_a)/3$. Then, for $n \geqslant n(\delta)$,

$$\log(T/K)/n \leqslant d(\mu_a + \delta, \mu^\star - \delta)\,.$$

We cut the sum in (3.26) at $n(\delta)$, so that

$$\begin{aligned} B &\leqslant n(\delta) - 1 + \sum_{n=n(\delta)}^{T} \mathbb{P}\big(d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant d(\mu_a + \delta, \mu^\star - \delta)\big) \\ &\leqslant \frac{\log(T/K)}{d(\mu_a + \delta, \mu^\star - \delta)} + \sum_{n=n(\delta)}^{T} \mathbb{P}\big(d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant d(\mu_a + \delta, \mu^\star - \delta)\big)\,. \end{aligned} \tag{3.27}$$

Using the inclusion

$$\big\{d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant d(\mu_a + \delta, \mu^\star - \delta)\big\} \subseteq \{\widehat{\mu}_{a,n} \geqslant \mu_a + \delta\}\,,$$

together with Inequality (3.47), we obtain that

$$\begin{aligned} \sum_{n=n(\delta)}^{T} \mathbb{P}\big(d(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant d(\mu_a + \delta, \mu^\star - \delta)\big) &\leqslant \sum_{n=n(\delta)}^{T} \mathbb{P}\big(\widehat{\mu}_{a,n} \geqslant \mu_a + \delta\big) \\ &\leqslant \sum_{n=1}^{\infty} e^{-n\delta^2/(2V)} = \frac{1}{e^{\delta^2/(2V)} - 1} \leqslant \frac{2V}{\delta^2}\,, \end{aligned}$$

and Equation (3.27) yields

$$B \leqslant \frac{\log(T)}{d(\mu_a + \delta, \mu^\star - \delta)} + \frac{2V}{\delta^2}\,. \tag{3.28}$$

Conclusion of the proof. It just remains to plug Inequalities (3.25) and (3.28) into Equation (3.21):

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T)}{d(\mu_a+\delta, \mu^\star-\delta)} + \frac{(2V)^2}{K\delta^4} + Te^{-\delta^2 T/(2KV)} + 1\,,$$

to obtain (3.6) and choose δ of order $1/\operatorname{loglog}(T)^{1/8}$ to obtain (3.7).

3.4 Refined Asymptotic Analysis for Bernoulli Rewards

In this section we present an asymptotic analysis of kl-UCB$^+$ algorithm with a second term of right order for Bernoulli rewards. Indeed, under a slightly stronger assumption on the strategy ψ (see Definition 2.4.2), one can proves that (see Theorem 2.4.3):

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{\log T}{\mathrm{kl}(\mu_a, \mu^\star)} - O\big(\log(\log T)\big)\,.$$

It appears that dividing by $N_a(t)$ in the exploration function allows also to catch this second order term. Indeed we have the following upper bound for the kl-UCB$^+$ algorithm.

Theorem 3.4.1. *If the horizon is such that* $T \geqslant \max\big(e^{4/\Delta_a^2}, \log(1/(1-\mu^\star))\big)$, *for all suboptimal* a,

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\log(T) - \operatorname{loglog}(T)}{\mathrm{kl}(\mu_a, \mu^\star)} + O(1)\,, \tag{3.29}$$

see Equation (3.42) *for an explicite expression of the term* $O(1)$.

A similar theorem holds for the kl-UCB^{++} algorithm since the extra factor K in the exploration function does not affect the asymptotic behaviour of the algorithm.

Sketch of arguments. As usual we use the decomposition (3.30) but we keep the index of the sub-optimal arm a in the term A. The main idea is that if $N_a(T)$ is too large, the index $U_a(t)$ should be close the mean μ_a. We can rule out this case with a second cutting, see Figure 3.1, and then just treat the case where $N_a(t)$ is of order $\log(T)$. This allow to take $\delta \sim 1/\log(T)$, the right order to obtain the second term in $-\operatorname{loglog}(T)$. For the term B we use a slightly modified version of Lemma 18 of Honda and Takemura [2015]. Usually we rewrite the term B, indexing it by the number of draws of arm a, as follows

$$\sum_{n=1}^{T} \mathbb{P}(U_{a,n} \geqslant \mu^\star - \delta)\,.$$

Then we split the sum in two. In order to exploit the fact that $\mathbb{P}(U_{a,n} \geqslant \mu^\star - \delta)$ is a deviation when $n \geqslant \log(T)/\mathrm{kl}(\mu_a, \mu^\star - \delta)$. The trick is to keep the fact that it is the same process inside each probability. And rather cut the sum at a well-chosen stopping time.

3.4.1 Proof of Theorem 3.4.1

Decomposition of $\mathbb{E}\big[N_a(T)\big]$. Let a be sub-optimal arm. As in the proof of Theorem 3.3.6 we can decompose $\mathbb{E}\big[N_a(T)\big]$ as follows:

$$\mathbb{E}\big[N_a(T)\big] \leqslant 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(U_a(t) \leqslant \mu^\star - \delta,\ A_{t+1} = a\big)}_{\text{A}} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\big(\mu^\star - \delta < U_a(t),\ A_{t+1} = a\big)}_{\text{B}} . \tag{3.30}$$

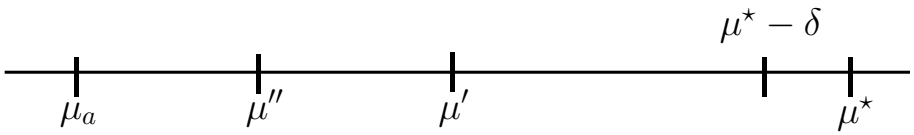

Figure 3.1: second cutting.

Term A. We cut again the event in B, according to Figure 3.1, at $\mu' := \mu^\star - \delta_0$, where $\delta_0 = (\mu^\star - \mu_a)/2$ is independent of δ

$$\begin{aligned} A &= \sum_{t=K}^{T-1} \mathbb{P}(U_a(t) \leqslant \mu',\ A_{t+1} = a) + \sum_{t=K}^{T-1} \mathbb{P}\big(\mu' \leqslant U_a(t) \leqslant \mu^\star - \delta,\ A_{t+1} = a\big) \\ &\leqslant T\,\mathbb{P}(\exists 1 \leqslant m \leqslant T, U_{a^*,m} \leqslant \mu') + \sum_{n=1}^{T} \mathbb{P}(\mu' \leqslant U_{a,n},\ \exists 1 \leqslant m \leqslant T,\ U_{a^*,m} \leqslant \mu^\star - \delta)\,. \end{aligned}$$

Let $\beta = \mathrm{kl}(\mu'', \mu')$ be the Kullback-Leibler divergence between $\mu'' = (\mu_a + \mu')/2$ and μ'. We cut the sum at the critical order of number of pulls: $\log(T)/\beta$, when the event $\{\mu' \leqslant U_{a,n}\}$ becomes atypical

$$\begin{aligned} A \leqslant T\,\mathbb{P}(\exists 1 \leqslant m \leqslant T, U_{a^*,m} \leqslant \mu') \quad &+ \sum_{1 \leqslant n \leqslant \log(T)/\beta} \mathbb{P}(\exists 1 \leqslant m \leqslant T,\ U_{a^*,m} \leqslant \mu^\star - \delta) \\ &+ \sum_{\log(T)/\beta < n} \mathbb{P}(\mu' \leqslant U_{a,n}) \\ \leqslant \underbrace{T\,\mathbb{P}(\exists 1 \leqslant m \leqslant T,\ U_{a^*,m} \leqslant \mu')}_{A_1} \quad &+ \underbrace{\frac{\log(T)}{\beta}\mathbb{P}(\exists 1 \leqslant m \leqslant T,\ U_{a^*,m} \leqslant \mu^\star - \delta)}_{A_2} \\ &+ \underbrace{\sum_{\log(T)/\beta < n} \mathbb{P}(\mu' \leqslant U_{a,n})}_{A_3}\,. \end{aligned}$$

Term A_1. Using the definition of $U_{a^\star,n}$ then Lemma (3.5.3) we get

$$\begin{aligned}\mathbb{P}(\exists 1 \leqslant m \leqslant T,\ U_{a^\star,m} \leqslant \mu') &\leqslant \mathbb{P}\big(\exists 1 \leqslant m \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^\star,m}, \mu^\star - \delta_0) \geqslant \log(T/m)/m\big) \\ &\leqslant \mathbb{P}\big(\exists 1 \leqslant m \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^\star,m}, \mu^\star) \geqslant \log(T/m)/m + 2\delta^2\big)\end{aligned}$$

where we write $\mathrm{kl}_+(p,q) = \mathrm{kl}(p,q)\mathbb{1}_{\{p\leqslant q\}}$. Now, thanks to the union bound and the not-maximal version of Lemma 3.5.5, we have

$$\begin{aligned}\mathbb{P}\big(\exists 1 \leqslant m \leqslant T,\ \mathrm{kl}_+(\widehat{\mu}_{a^\star,m}, \mu^\star) \geqslant \log(T/m)/m + 2\delta^2\big) &\leqslant \sum_{m=1}^{+\infty} \frac{m}{T} e^{-2\delta_0^2 m} \\ &= \frac{e^{2\delta_0^2}}{T\big(e^{2\delta_0^2}-1\big)^2} \leqslant \frac{e^2}{4T\delta_0^4}\,.\end{aligned}$$

Chaining these inequalities leads to the upper bound

$$A_1 \leqslant \frac{e^2}{4\delta_0^4} = \frac{4e^2}{\Delta_a^4}, \tag{3.31}$$

where we used that by definition $\delta_0 = (\mu^\star - \mu_a)/2$.

Term A_2. We can proceed exactly in the same way as above, replacing δ_0 by δ, to show that

$$A_2 \leqslant \frac{\log(T)}{\beta}\frac{e^2}{4T\delta^4} \leqslant \frac{\log(T)}{T}\frac{2e^2}{\delta^4\Delta_a^2}\,, \tag{3.32}$$

where we used in the second inequality $\beta = \mathrm{kl}(\mu'', \mu') \geqslant 2(\mu''-\mu')^2 = \Delta_a^2/8$.

Term A_3. For this term we use the fact that for n large enough the upper-confidence bound $U_{a,n}$ is close to the mean μ_a. Indeed by definition of the index and for $n > \log(T)/\beta$, we have

$$\begin{aligned}\mathbb{P}(U_{a,n} \geqslant \mu') &\leqslant \mathbb{P}\big(\mathrm{kl}_+(\widehat{\mu}_{a,n}, \mu') \leqslant \log(T/n)/n\big) \\ &\leqslant \mathbb{P}\big(\mathrm{kl}_+(\widehat{\mu}_{a,n}, \mu') \leqslant \mathrm{kl}(\mu'', \mu')\big)\,.\end{aligned}$$

Now since we have the inclusion $\{\mathrm{kl}_+(\widehat{\mu}_{a,n}, \mu') \leqslant \mathrm{kl}(\mu'', \mu')\} \subset \{\widehat{\mu}_{a,n} \leqslant \mu''\}$, the Hoeffding Inequality (3.47) entails

$$\begin{aligned}A_3 &\leqslant \sum_{n=1}^{T} e^{-2(\mu''-\mu_a)^2} \leqslant \frac{1}{e^{2(\mu''-\mu_a)^2}-1} \\ &\leqslant \frac{8}{\Delta_a^2}\,.\end{aligned} \tag{3.33}$$

Putting Equation (3.31), (3.32) and (3.33) leads to the upper-bound

$$A \leqslant \frac{\log(T)}{T}\frac{2e^2}{\delta^4\Delta_a^2} + \frac{4e^2+8}{\Delta_a^4}\,. \tag{3.34}$$

Term B We index the sum by the number of draws of arm a and use the definition of the index to get

$$\begin{aligned} B &\leqslant \sum_{n=1}^{T} \mathbb{P}(U_{a,n} \geqslant \mu^\star - \delta) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}\big(\mathrm{kl}(\widehat{\mu}_{a,n}, \mu^\star - \delta) \leqslant \log(T/n)/n\big)\,. \end{aligned} \tag{3.35}$$

Let λ_a be the negative real such that $\mathrm{kl}(\mu_a, \mu^\star - \delta) = \lambda_a \mu_a - \varphi_{\mu^\star - \delta}(\lambda_a)$. Where $\varphi_{\mu^\star-\delta}$ is the log-partition function of $\mathrm{Ber}(\mu^\star - \delta)$, i.e.

$$\varphi_{\mu^\star-\delta}(\lambda) = \log\left(e^{\lambda}(\mu^\star - \delta) + 1 - (\mu^\star - \delta)\right).$$

We denote by $Z_{a,k}$ the random variable

$$Z_{a,k} := \lambda_a X_{k,a} - \varphi_{\mu^\star-\delta}(\lambda_a)\,, \tag{3.36}$$

and introduce the boundary crossing stopping time

$$\tau = \inf\left\{ n : \sum_{k=1}^{n} Z_{a,k} > \frac{\log(T/n)}{n} \right\}.$$

Thanks to the variational formula of the Kullback-Leibler divergence and by definition of τ, Equation (3.35) leads to

$$\begin{aligned} B &\leqslant \sum_{n=1}^{T} \mathbb{P}\left(\sum_{k=1}^{n} Z_{a,k} \leqslant \log(T/n)/n \right) \\ &\leqslant \mathbb{E}[\tau - 1] + \mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{ \sum_{k=1}^{n} Z_{a,k} \leqslant \log(T/n)/n \right\}} \right] \end{aligned} \tag{3.37}$$

On one hand, thanks to Lemma 4.6.1, we can upper bound the expectation of the stopping time

$$\mathbb{E}[\tau] \leqslant \frac{W(TM) + M + \log(2)}{\mathrm{kl}(\mu_a, \mu^\star - \delta)}\,, \tag{3.38}$$

where we denote the Lambert function by W and by M an upper-bound on the random variables $Z_{a,k}$:

$$Z_{a,k} \leqslant -\varphi_{\mu^\star-\delta}(\lambda_a) = \log\left(\frac{1}{e^{\lambda_a}(\mu^\star - \mu) + 1 - (\mu^\star - \delta)} \right) \leqslant M := \log\left(\frac{1}{1 - \mu^\star} \right). \tag{3.39}$$

On the other hand, to bound the second expectation of (3.37), note that for $n > \tau$, we have

$$\begin{aligned} \sum_{k=1}^{n} Z_{a,k} &\geqslant \log(T/\tau)/\tau + \sum_{k=\tau+1}^{n} Z_{a,k} \\ &\geqslant \log(T/n)/n + \sum_{k=\tau+1}^{n} Z_{a,k}\,. \end{aligned}$$

Thus, using this inequality and conditioning by τ, one obtains

$$\mathbb{E}\Bigg[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{k=1}^{n} Z_{a,k} \leqslant \log(T/n)/n\right\}}\Bigg] \leqslant \mathbb{E}\Bigg[\sum_{n=\tau+1}^{T} \mathbb{P}\Bigg(\sum_{k=1}^{n} Z_{a,k} \leqslant 0 \Bigg| \tau\Bigg)\Bigg],$$

Now, we use Lemma 3.5.6 then Pinsker Inequality (3.1) to control the probability inside the expectation:

$$\mathbb{P}\Bigg(\sum_{k=\tau+1}^{n} Z_{a,k} \leqslant 0 \Bigg| \tau\Bigg) \leqslant e^{-(n-\tau)(1-\mu^\star)^{5/2}\mathrm{kl}(\mu_a,\mu^\star-\delta)/18} \leqslant e^{-(n-\tau)(1-\mu^\star)^{5/2}\Delta_a^2/36}.$$

Summing these inequalities and re-indexing, we ends up with

$$\begin{aligned}\mathbb{E}\Bigg[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{k=1}^{n} Z_{a,k} \leqslant \log(T/n)/n\right\}}\Bigg] &\leqslant \mathbb{E}\Bigg[\sum_{n=\tau+1}^{T} e^{-(n-\tau)(1-\mu^\star)^{5/2}\Delta^2/36}\Bigg] \\ &\leqslant \frac{1}{1-e^{-(1-\mu^\star)^{5/2}\Delta_a^2/36}} \leqslant \frac{36e}{(1-\mu^\star)^{5/2}\Delta_a^2}. \qquad (3.40)\end{aligned}$$

Putting Inequality (3.37), (3.38) and (3.40) together we obtain

$$B \leqslant \frac{W(TM)+M+\log(2)}{\mathrm{kl}(\mu_a,\mu^\star-\delta)} + \frac{36e}{(1-\mu^\star)^{5/2}\Delta_a^2}. \qquad (3.41)$$

Conclusion Combining Inequality (3.34) and (3.41) leads to

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant 1 + \frac{W(TM)+M+\log(2)}{\mathrm{kl}(\mu_a,\mu^\star-\delta)} + \frac{36e}{(1-\mu^\star)^{5/2}\Delta_a^2} \\ \frac{\log(T)}{T}\frac{2e^2}{\delta^4\Delta_a^2} + \frac{4e^2+8}{\Delta_a^4}. \qquad (3.42)\end{aligned}$$

An inequality on the Lambert function from Lemma 3.5.1, Inequality (3.45) on the Kullback-Leibler divergence and the choice $\delta = 1/\log(T)$ allow us to conclude.

3.5 Elements of Proofs

3.5.1 Lambert function

We present here some inequalities on the Lambert function from Hoorfar and Hassani [2008b].

Lemma 3.5.1. *For all* $x \geqslant e$,

$$\log(x) - \operatorname{loglog}(x) \leqslant W(x) \leqslant \log(x) - \operatorname{loglog}(x) + \log(1 + e^{-1}) \,.$$

3.5.2 Inequalities involving the Kullback-Leibler Divergence.

A useful representation of the Kullback-Leibler divergence is the variational formula.

Lemma 3.5.2. *For all* $(\mu, \mu') \in I^2$,

$$d(\mu, \mu') = \sup_{\lambda \in \mathbb{R}} \lambda\mu - \varphi_\mu(\lambda) \,, \tag{3.43}$$

where φ_μ *is the log-partition function of the distribution* $\nu_{b^\star(\mu)}$.

An inequality that is a consequence of generalized law of cosines.

Lemma 3.5.3. *For all* $(\mu', \mu) \in I^2$ *and* $\delta \geqslant 0$ *such that* $\mu' \leqslant \mu - \delta \in I$,

$$d(\mu', \mu - \delta) + \delta^2/(2V) \leqslant d(\mu', \mu) \,. \tag{3.44}$$

Proof. Thanks to the generalized law of cosines and $\mu' \leqslant \mu - \delta \leqslant \mu$, we have

$$\begin{aligned} d(\mu', \mu - \delta) + d(\mu - \delta, \mu) &= d(\mu', \mu) + (\mu' - \mu + \delta)(b'^{-1}(\mu) - b'^{-1}(\mu - \delta)) \\ &\leqslant d(\mu', \mu) \,. \end{aligned}$$

The Pinsker Inequality (3.1): $d(\mu - \delta, \mu) \geqslant \delta^2/(2V)$ allow us to conclude. □

A reverse inequality holds for Kullback-Leibler divergence between Bernoulli. It is a simple consequence of Lemma 4.5.4.

Lemma 3.5.4. *For all* $(\mu', \mu) \in (0,1)^2$ *and* $0 < \delta < \mu$,

$$\operatorname{kl}(\mu', \mu) \leqslant \operatorname{kl}(\mu', \mu) + \frac{\varepsilon}{1 - \mu} \,. \tag{3.45}$$

3.5.3 Deviation-concentration inequalities

We regroup in this section some deviation inequalities used in the proofs of this chapter.

Lemma 3.5.5. *(Maximal Inequality) Let N and M be two real numbers in $\mathbb{R}^+ \times \overline{\mathbb{R}^+}$, let γ be a real number in $\mathbb{R}^{+*}$, and let $\widehat{\mu}_n$ be the empirical mean of n random variables i.i.d. according to the distribution $\nu_{b'^{-1}(\mu)}$. Then*

$$\mathbb{P}\big(\exists N \leqslant n \leqslant M, \quad d_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big) \leqslant e^{-N\gamma}\,. \tag{3.46}$$

Proof. If $\gamma > d_+(\inf(I), \mu)$ or $\widehat{\mu}_n \geqslant \mu$ the Inequality (3.46) is trivial. Else, thanks to the variational formula of the Kullback-leibler divergence, there exists two real numbers $z < \mu$ and $\lambda < 0$ such that

$$\gamma = d(z, \mu) = \lambda z - \varphi_\mu(\lambda)\,,$$

where φ_μ denotes the the log-moment generating function of $\nu_{b'^{-1}(\mu)}$. Since on the event $\big\{\exists N \leqslant n \leqslant M, \quad d_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big\}$ one has at the same time

$$\widehat{\mu}_n \leqslant \mu\,, \quad \lambda\widehat{\mu}_n - \varphi_\mu(\lambda) \geqslant \lambda z - \varphi_\mu(\lambda) = \gamma \quad \text{and} \quad \lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda) \geqslant N\gamma\,,$$

we can write that

$$\begin{aligned}\mathbb{P}\big(\exists N \leqslant n \leqslant M, \quad d_+(\widehat{\mu}_n, \mu) \geqslant \gamma\big) &\leqslant \mathbb{P}\big(\exists N \leqslant n \leqslant M, \quad \lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda) \geqslant N\gamma\big)\\ &\leqslant \exp(-N\gamma)\,,\end{aligned}$$

by Doob's maximal inequality for the exponential martingale $\exp\big(\lambda n\widehat{\mu}_n - n\varphi_\mu(\lambda)\big)$. □

As a simple consequence of this Lemma 3.5.5 and Inequality (3.1), it holds that:

$$\text{for every } x \leqslant \mu, \qquad \mathbb{P}(\exists N \leqslant n \leqslant M, \quad \widehat{\mu}_n \leqslant x) \leqslant e^{-N(x-\mu)^2/(2V)}\,, \tag{3.47}$$

$$\text{for every } x \geqslant \mu, \qquad \mathbb{P}(\exists N \leqslant n \leqslant M, \quad \widehat{\mu}_n \geqslant x) \leqslant e^{-N(x-\mu)^2/(2V)}\,. \tag{3.48}$$

We can integrate these inequalities to obtain bound on the following expectation,

$$\text{for every } \delta \geqslant 0, \qquad \mathbb{E}\big[\max_{N\leqslant n\leqslant M}(\mu - \widehat{\mu}_n - \delta)^+\big] \leqslant \sqrt{\frac{\pi V}{2N}}e^{-N\delta^2/(2V)}\,, \tag{3.49}$$

and for n fix, we get

$$\text{for every } \delta \geqslant 0, \qquad \mathbb{E}\big[(\widehat{\mu}_n - \mu - \delta)^+\big] \leqslant \sqrt{\frac{\pi V}{2n}}e^{-n\delta^2/(2V)}\,. \tag{3.50}$$

Proof of Inequality 3.49. Integrating the Inequality (3.47), leads to

$$\begin{aligned}E\big[\max_{N\leqslant n\leqslant M}(\mu - \widehat{\mu}_n - \delta)^+\big] &\leqslant \int_0^{+\infty} \mathbb{P}(\exists N \leqslant n \leqslant M, \quad \widehat{\mu}_n \leqslant \mu - \delta - u)\mathrm{d}u\\ &\leqslant \int_0^{+\infty} e^{-N(\delta+u)^2/(2V)}\mathrm{d}u\\ &\leqslant e^{-N\delta^2/(2V)}\int_0^{+\infty} e^{-Nu^2/(2V)}\mathrm{d}u = \sqrt{\frac{\pi V}{2N}}e^{-N\delta^2/(2V)}\,.\end{aligned}$$

□

Lemma 3.5.6. *Fix $0 < \mu < \widetilde{\mu} < 1$. Let the random variables $(Z_k)_{1\leqslant k\leqslant n}$ be such that $Z_k = \lambda X_k - \varphi_{\widetilde{\mu}}(\lambda)$ where $(X_k)_{1\leqslant k\leqslant n}$ are i.i.d. according to the distribution $\mathrm{Ber}(\mu)$, $\varphi_{\widetilde{\mu}}$ is the log-moment-generating function of $\mathrm{Ber}(\mu)$ and λ is such that*

$$\mathrm{kl}(\mu, \widetilde{\mu}) = \lambda\mu - \varphi_{\widetilde{\mu}}(\lambda)\,.$$

Then for all $-\mathrm{kl}(\mu, \widetilde{\mu}) < u < \mathrm{kl}(\mu, \widetilde{\mu})$, it holds

$$\mathbb{P}\Big(\sum_{k=1}^{n} Z_k \leqslant u\Big) \leqslant \frac{(1-\mu)^{5/2}}{18}\big(\mathrm{kl}(\mu, \widetilde{\mu}) - u\big)^2 \tag{3.51}$$

Proof. Since $\mathbb{E}[Z_1] = \mathrm{kl}(\mu, \widetilde{\mu})$, thanks to the Cramèr-Chernoff bound, we have

$$\mathbb{P}\Big(\sum_{k=1}^{n} Z_k \leqslant \mathrm{kl}(\mu, \widetilde{\mu}) - u\Big) \leqslant \exp\big(-n \sup_{x\leqslant 0} xu - \psi(x)\big)\,,$$

where ψ is log-moment-generating function of Z_1. It remains to prove that

$$\sup_{x\leqslant 0} xu - \psi(x) \geqslant \frac{(1-\mu)^{5/2}}{18}\big(\mathrm{kl}(\mu, \widetilde{\mu}) - u\big)^2\,.$$

A Taylor inequality for the function ψ, entails that for all $x \in [-1/2, 0]$

$$\begin{aligned}\psi(x) &\leqslant \psi(0) + \psi'(0)x + C\frac{x^2}{2}\\ &= x\mathrm{kl}(\mu, \widetilde{\mu}) + C\frac{x^2}{2}\,,\end{aligned}$$

where

$$C := \sup_{x\in[-1/2,0]} \psi''(x)\,.$$

To upper bound C, we remark that

$$\psi''(x) = \mathbb{E}\left[Z_1^2 \frac{e^{xZ_1}}{\mathbb{E}[e^{xZ_1}]}\right] - E\left[Z_1 \frac{e^{xZ_1}}{\mathbb{E}[e^{xZ_1}]}\right]^2 \tag{3.52}$$

$$\leqslant \frac{\mathbb{E}\left[Z_1^2 e^{xZ_1}\right]}{\mathbb{E}[e^{xZ_1}]}\,. \tag{3.53}$$

Thanks to the Inequality (3.39), we have for $-1/2 \leqslant x \leqslant 0$

$$\mathbb{E}\big[e^{xZ_1}\big] \geqslant \big(1/(1-\widetilde{\mu})\big)^x \geqslant \sqrt{1-\widetilde{\mu}}\,. \tag{3.54}$$

Then for all $z \in \Big(-\infty, \log\big(1/(1-\mu)\big)\Big]$ and $-1/2 \leqslant x \leqslant 0$ it holds

$$z^2 e^{xz} \leqslant 8e^{-z} + \frac{1}{(1-\mu)^2}\,. \tag{3.55}$$

Indeed, if $z \geqslant 0$ it immediately follows that

$$z^2 e^{xz} \leqslant z^2 \leqslant \log\left(\frac{1}{1-\mu}\right)^2 \leqslant \frac{1}{(1-\mu)^2}\,,$$

and if $z \leqslant 0$, using $z^2 \leqslant 8e^{-z/2}$ in this case, we obtain

$$z^2 e^{xz} \leqslant 8e^{(x-1/2)z} \leqslant 8e^{-z}\,.$$

Combining (3.55) and the fact that

$$\mathbb{E}[e^{-Z_1}] = \mathbb{E}_{\widetilde{\mu}}[1] = 1$$

one obtains

$$E\big[Z_1^2 e^{xZ_1}\big] \leqslant 8E\big[e^{-Z_1}\big] + \frac{1}{(1-\widetilde{\mu})^2} \leqslant \frac{9}{(1-\widetilde{\mu})^2}\,. \tag{3.56}$$

Putting all together, i.e. equations (3.53), (3.54) and (3.56), we get

$$C \leqslant \frac{9}{(1-\widetilde{\mu})^{5/2}}\,,$$

and therefore, for all $x \in [-1/2, 0]$

$$\psi(x) \leqslant \mathrm{kl}(\mu, \widetilde{\mu})x + \frac{9}{(1-\widetilde{\mu})^{5/2}}\frac{x^2}{2}\,.$$

Thus, optimizing in x, we obtain

$$\begin{aligned}\sup_{x\leqslant 0} xu - \psi(x) &\geqslant \sup_{-1/2\leqslant x<0} x\big(u - \mathrm{kl}(\mu,\widetilde{\mu})\big) - \frac{9}{(1-\widetilde{\mu})^{5/2}}\frac{x^2}{2} \\ &= \frac{(1-\mu)^{5/2}}{18}\big(\mathrm{kl}(\mu,\widetilde{\mu}) - u\big)^2\,,\end{aligned}$$

since the maximum is attained at

$$\begin{aligned}x = -\big(\mathrm{kl}(\mu,\widetilde{\mu}) - u\big)\,\frac{(1-\widetilde{\mu})^{5/2}}{9} &\geqslant -2\log\frac{1}{1-\widetilde{\mu}}\,\frac{(1-\widetilde{\mu})^{5/2}}{9} \\ &\geqslant -\frac{2}{9}(1-\widetilde{\mu})^{3/2} \geqslant -\frac{1}{2}\end{aligned}$$

thanks to the assumption $-\mathrm{kl}(\mu,\widetilde{\mu}) < u < \mathrm{kl}(\mu,\widetilde{\mu})$. □

Chapter 4

KL-UCB Algorithms for Bounded Rewards

In collaboration with Aurélien Garivier, Hédi Hadiji and Gilles Stoltz.

Contents

4.1 Introduction and brief literature review

In this Chapter we extend the results of the previous chapter to bandit problems with bounded rewards. Precisely these of the Bernoulli bandit problems. Roughly speaking, we move from a parametric problem to a non-parametric one. In the early 2000s, the much noticed contributions of Auer et al. [2002a] and Auer et al. [2002b] promoted three important ideas.

1. First, a bandit strategy should not address only specific statistical models as in Chapter 3, but general and non-parametric families of probability distributions, e.g., bounded distributions.

2. Second, the regret analysis should not only be asymptotic, but should provide finite-time bounds.

3. Third, a good bandit strategy should be competitive with respect to two concurrent notions of optimality: distribution-dependent optimality (it should reach the asymptotic lower bound of Lai and Robbins, cf Theorem 2.2.5) and minimax optimality (the maximal regret over all considered probability distributions should be of the optimal order $\sqrt{KT}$, cf Inequality 2.3).

Initiated by Honda and Takemura for the IMED algorithm (see Honda and Takemura, 2015 and references to earlier works of the authors therein) and followed by Cappé et al. [2013] for the KL-UCB algorithm, the use of the empirical likelihood method for the construction of the upper confidence bounds was proved to be optimal as far as distribution-dependent bounds are concerned. The analysis for IMED was led for all (semi-)bounded distributions, while the analysis for KL-UCB was only sucessfully achieved in some classes of distributions (e.g., bounded distributions with finite supports). A contribution in passing of the present chapter is to also provide optimal distribution-dependent bounds for KL-UCB for families of bounded distributions.

On the other hand, classical UCB strategies were proved not to enjoy distribution-free optimal regret bounds. A modified strategy named MOSS was proposed by Audibert and Bubeck [2009] to address this issue: minimax optimality (for bounded distributions) was proved, but distribution-dependent optimality was then not considered. It took a few more years before Ménard and Garivier [2017] (cf. Chapter 3) and Lattimore [2018] (see also Lattimore [2016]) proved that, in simple parametric settings, a strategy can enjoy, at the same time, regret bounds that are optimal both from a distribution-dependent and a distribution-free viewpoints.

Main contributions. In this work, we generalize the latter bi-optimality result of Chapter 3 (for the Bernoulli bandit problem) to the non-parametric class of distributions with bounded support, say, $[0, 1]$. Namely, we propose the KL-UCB-switch algorithm, a bandit strategy belonging to the family of upper-confidence-bounds strategies. We prove that it is simultaneously optimal from a distribution-free viewpoint (Theorem 4.2.1)

and from a distribution-dependent viewpoint in the considered class of distributions (Theorem 4.2.2).

We go one step further by providing, as Honda and Takemura [2015] already achieved for IMED, a second-order term of the optimal order $-\log(\log(T))$ in the distribution-dependent bound (Theorem 4.2.3). This explains from a theoretical viewpoint why simulations consistently show strategies having a regret smaller than the main term of the lower bound of Lai and Robbins [1985]. Note that, to the best of our knowledge, IMED is not proved to enjoy an optimal distribution-free regret bound; only a distribution-dependent regret analysis was provided for it.

Beyond these results, we took special care of the clarity and simplicity of all the proofs, and all our bounds are finite time, with closed-form expressions. In particular, we provide for the first time an elementary analysis of performance of the KL-UCB algorithm on the class of all distributions over a bounded interval. The study of KL-UCB in Cappé et al. [2013] indeed remained somewhat intricate and limited to finitely supported distributions. Furthermore, our simplified analysis allowed us to derive similar optimality results for the anytime version of this new algorithm, with little if no additional effort (see Theorems 4.2.4 and 4.2.5).

Organization of the chapter. Section 4.2 contains the presentation of the KL-UCB-switch algorithm, the precise statement of the aforementioned theorems, and corresponding results for an anytime version of the KL-UCB-switch algorithm. Section 4.3 discusses some numerical experiments comparing the performance of the KL-UCB-switch algorithm to competitors like IMED or KL-UCB. Section 4.5 contains the statements and the proofs of several results that were already known before, but for which we sometimes propose a simpler derivation. All technical results needed in this chapter are thus stated and proved from scratch (e.g., on the $\mathcal{K}_{\inf}$ quantity that is central to the analysis of IMED and KL-UCB, and on the analysis of the performance of MOSS), which makes our submission fully self-contained. These known results are used as building blocks in Section 4.4, where the main results of this chapter are proved, up to some more sophisticated bound whose analysis is detailed in Section 4.6. Technical arguments are deferred to the appendices.

4.2 Setting and statement of the main results

We consider the simplest case of a stochastic bandit problem, with finitely many arms indexed by $a \in \{1, \dots, K\}$. Each of these arms is associated with an unknown probability distribution ν_a over $[0,1]$. We call $\underline{\nu} = (\nu_1, \dots, \nu_K)$ a bandit problem over $[0,1]$. At each round $t \geqslant 1$, the player pulls the arm A_t and gets a real-valued reward Y_t drawn independently at random according to the distribution ν_{A_t}. This reward is the only piece of information available to the player.

A typical measure of the performance of a strategy is given by its *regret*. To recall its definition, we denote by $\mathrm{E}(\nu_a) = \mu_a$ the expected payoff of arm a and by Δ_a its gap

to an optimal arm:

$$\mu^\star = \max_{a=1,\dots,K} \mu_a \qquad \text{and} \qquad \Delta_a = \mu^\star - \mu_a \,.$$

Arms a such that $\Delta_a > 0$ are called suboptimal arms. The expected regret of a strategy equals

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T Y_t\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T \mu_{A_t}\right] = \sum_{a=1}^K \Delta_a \, \mathbb{E}\big[N_a(T)\big] \quad \text{where} \quad N_a(T) = \sum_{t=1}^T \mathbb{1}_{\{A_t=a\}}.$$

The first equality above follows from the tower rule. To control the expected regret, it is thus sufficient to control the $\mathbb{E}\big[N_a(T)\big]$ quantities for suboptimal arms a.

Reminder of the existing lower bounds. The distribution-free lower bound of Auer et al. [2002b] states that for all strategies, for all $T \geqslant 1$ and all $K \geqslant 2$,

$$\sup_{\underline{\nu}} R_T \geqslant \frac{1}{20} \min\Big\{\sqrt{KT}, \, T\Big\}\,, \tag{4.1}$$

where the supremum is taken over all bandit problems $\underline{\nu}$ over $[0,1]$ (see Inequality 2.3 of Chapter 2 and its proof in Section 2.5.2).

We denote by $\mathcal{P}[0,1]$ the set of all distributions over $[0,1]$. The key quantity in stating distribution-dependent lower bounds is based on KL, the Kullback-Leibler divergence between two probability distributions. For $\nu_a \in \mathcal{P}[0,1]$ and $x \in [0,1]$,

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) : \quad \nu'_a \in \mathcal{P}[0,1] \quad \text{and} \quad \mathrm{E}(\nu'_a) > x\Big\}\,,$$

where $\mathrm{E}(\nu'_a)$ denotes the expectation of the distribution ν'_a and where by convention, the infimum of the empty set equals $+\infty$. As essentially proved by Lai and Robbins [1985] and Burnetas and Katehakis [1996]—see also Theorem 2.2.5—, for any "reasonable" strategy, for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\big[N_a(T)\big]}{\log T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,. \tag{4.2}$$

By "reasonable" strategy, we mean a strategy that is uniformly fast convergent on $\mathcal{P}[0,1]$, that is, such that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\forall\, \alpha > 0, \qquad \mathbb{E}\big[N_a(T)\big] = o(T^\alpha)\,.$$

For uniformly super-fast convergent strategies, that is, strategies for which there exists a constant C such for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\frac{\mathbb{E}\big[N_a(T)\big]}{\log T} \leqslant \frac{C}{\Delta_a^2}\,,$$

the lower bound above can be strengthened into: for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - \Omega\big(\log(\log T)\big)\,, \tag{4.3}$$

see Theorem 2.4.3 for an exact statement and its proof. This order of magnitude $-\log(\log T)$ for the second-order term in the regret bound is optimal, as follows from the upper bound exhibited by Honda and Takemura [2015, Theorem 5].

4.2.1 The KL-UCB-switch algorithm

Algorithm 9: Generic index policy

Inputs: index functions U_a
Initialization: Play each arm $a = 1, \dots, K$ once and compute the $U_a(K)$
for $t = K+1, \dots, T$ **do**
  Pull an arm $A_t \in \arg\max_{a=1,\dots,K} U_a(t-1)$
  Get a reward Y_t drawn independently at random according to ν_{A_t}
end for

For any index policy as described above, we have $N_a(t) \geqslant 1$ for all arms a and $t \geqslant K$ and may thus define, respectively, the empirical distribution of the rewards associated with arm a up to round t included and their empirical mean:

$$\widehat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{1}_{\{A_s = a\}} \qquad \text{and} \qquad \widehat{\mu}_a(t) = \mathrm{E}\big[\widehat{\nu}_a(t)\big] = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s = a\}}\,,$$

where δ_y denotes the Dirac point-mass distribution at $y \in [0,1]$.

The MOSS algorithm (see Audibert and Bubeck [2009]) uses the index functions

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\!\left(\frac{T}{KN_a(t)}\right)}\,, \tag{4.4}$$

where $\log_+$ denotes the nonnegative part of the natural logarithm, $\log_+ = \max\{\log, 0\}$.

We also consider a slight variation of the KL-UCB algorithm (see Cappé et al. 2013), which we call KL-UCB$^+$ and which relies on the index functions

$$U_a^{\mathrm{KL}}(t) \stackrel{\mathrm{def}}{=} \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)} \log_+\!\left(\frac{T}{KN_a(t)}\right) \right\}. \tag{4.5}$$

We introduce a new algorithm KL-UCB-switch. The novelty here is that this algorithm switches from the KL-UCB-type index to the MOSS index once it has pulled an arm more than $f(T,K)$ times. In the sequel we will take $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$. More precisely, we define the index functions

$$U_a(t) = \begin{cases} U_a^{\mathrm{KL}}(t) & \text{if } N_a(t) \leqslant f(T,K) \\ U_a^{\mathrm{M}}(t) & \text{if } N_a(t) > f(T,K) \end{cases}$$

4.2.2 Optimal distribution-dependent and distribution-free regret bounds (known horizon T)

We first consider a fixed and beforehand-known value of T. The proofs of the theorems below are provided in Section 4.4.

Theorem 4.2.1 (Distribution-free bound)**.** *Given $T \geqslant 1$, the regret of the KL-UCB-switch algorithm, tuned with the knowledge of T and the switch function $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$, is uniformly bounded over all bandit problems $\underline{\nu}$ over $[0,1]$ by*

$$R_T \leqslant (K-1) + 25\sqrt{KT}\,,$$

KL-UCB-switch thus enjoys a distribution-free regret bound of optimal order $\sqrt{KT}$, see (4.1). That was already the case for the MOSS strategy by Audibert and Bubeck [2009].

Theorem 4.2.2 (Distribution-dependent bound)**.** *Given $T \geqslant 1$, the KL-UCB-switch algorithm, tuned with the knowledge of T and the switch function $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$, ensures that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + O_T\big((\log T)^{2/3}\big)\,,$$

where a finite-time, closed-formed expression of the $O_T\big((\log T)^{2/3}\big)$ term is the sum of the bounds (4.13) and (4.16) for the choice $\delta = (\log T)^{-1/3}$.

By considering the exact same algorithm but by following a more sophisticated proof we may in fact get a stronger result.

Theorem 4.2.3 (Distribution-dependent bound with a second-order term)**.** *We actually have*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\log T - \log\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + O_T(1)\,,$$

where a finite-time, closed-formed expression of the $O_T(1)$ term is the sum of the bounds (4.13) and (4.39) for the choice $\delta = T^{-1/8}$.

KL-UCB-switch thus enjoys a distribution-distribution regret bounds of optimal orders, see (4.2) and (4.3). That was already the case for the IMED strategy by Honda and Takemura [2015] on the model $\mathcal{P}[0,1]$. The KL-UCB algorithm studied, e.g., by Cappé et al. [2013], only enjoyed optimal regret bounds for more limited models; for instance, for distributions over $[0,1]$ with finite support. In the analysis of KL-UCB-switch we actually provide in passing an analysis of KL-UCB for the model $\mathcal{P}[0,1]$ of all distributions over $[0,1]$.

4.2.3 Adaptation to the horizon T (an anytime version of KL-UCB-switch)

A standard doubling trick fails to provide a meta-strategy that would not require the knowledge of T and have optimal $O\big(\sqrt{KT}\big)$ and $\big(1+o(1)\big)(\log T)/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ bounds. Indeed, there are first, two different rates, $\sqrt{T}$ and $\log T$, to accommodate simultaneously and each would require different regime lengths, e.g., 2^r and 2^{2^r}, respectively, and second, any doubling trick on the distribution-dependent bound would result in an additional multiplicative constant in front of the $1/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ factor. This is why a dedicated anytime version of our algorithm is needed.

For technical reasons, it was useful in our proof to perform some additional exploration, which deteriorates the second-order terms in the regret bound. Indeed, we define the augmented exploration function

$$\varphi(x) = \log_+\big(x(1+\log_+^2 x)\big) \tag{4.6}$$

and the corresponding anytime index

$$U_a^{\mathrm{ANY}}(t) = \begin{cases} \sup\left\{ \mu\in[0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t),\mu\big) \leqslant \dfrac{1}{N_a(t)}\,\varphi\!\left(\dfrac{t}{KN_a(t)}\right)\right\} & \text{if } N_a(t)\leqslant f(t,K) \\ \widehat{\mu}_a(t) + \sqrt{\dfrac{1}{2N_a(t)}\,\varphi\!\left(\dfrac{t}{KN_a(t)}\right)} & \text{if } N_a(t) > f(t,K) \end{cases}$$

Theorem 4.2.4 (Anytime distribution-free bound)**.** *The regret of the anytime version of KL-UCB-switch algorithm above, tuned with the switch function $f(t,K)=\lfloor (t/K)^{1/5}\rfloor$, is uniformly bounded over all bandit problems $\underline{\nu}$ over $[0,1]$ as follows: for all $T\geqslant 1$,*

$$R_T \leqslant (K-1) + 46\sqrt{KT}$$

Theorem 4.2.5 (Anytime distribution-dependent bound)**.** *The anytime version of KL-UCB-switch algorithm above, tuned with the switch function $f(t,K)=\lfloor (t/K)^{1/5}\rfloor$, ensures that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}$$

We provide the proofs of the two theorems in Appendix 4.7.3. The distribution-free analysis is essentially the same as in the case of a known horizon, although the additional exploration required an adaptation of most of the calculations. Note also that the simulations detailed below suggest that all anytime variants of the KL-UCB algorithms (KL-UCB-switch included) behave better without the additional exploration required, i.e., with $\log_+$ as the exploration function.

4.3 Numerical experiments

We start by describing the algorithms used in all experiments (and their parameters): of course, KL-UCB-switch, KL-UCB, and MOSS, as described in Section 4.2.1. We actually consider their anytime versions, see Sections 4.2.3, and to do use, resort for all three of them to the exploration function $\log\big(t/(KN_a(t))\big)$. We also consider the IMED strategy of Honda and Takemura [2015].

For KL-UCB-switch we actually consider a bit more aggressive switch $f(t,K) = \lfloor t/K \rfloor^{8/9}$ than in our theoretical analysis; while our choice $f(t,K) = \lfloor t/K \rfloor^{1/5}$ appeared most naturally in the proofs, many other choices were possible at the cost of higher constants in one of the regret bounds.

Distribution-dependent bounds. We compare in Figure 4.1 the distribution-dependent behaviors of the algorithms. For the two scenarios with truncated exponential or Gaussian rewards we also consider the appropriate version of the kl-UCB algorithm for one-parameter exponential family (see Cappé et al., 2013), with the same exploration function as for the other algorithms; we call these algorithms kl-UCB-exp or kl-UCB-Gauss, respectively. The parameters of the middle and right scenarios were chosen in a way that, even with the truncation, the kl-UCB algorithms have a significantly better performance than the other algorithms. (This is the case because they are able to exploit the form of the underlying distributions.) Note that the kl-UCB-gauss algorithm reduces to the MOSS algorithm with the constant $2\sigma^2$ instead of $1/2$.

As expected the regret of KL-UCB-switch is an interpolation between the one of MOSS and of KL-UCB.

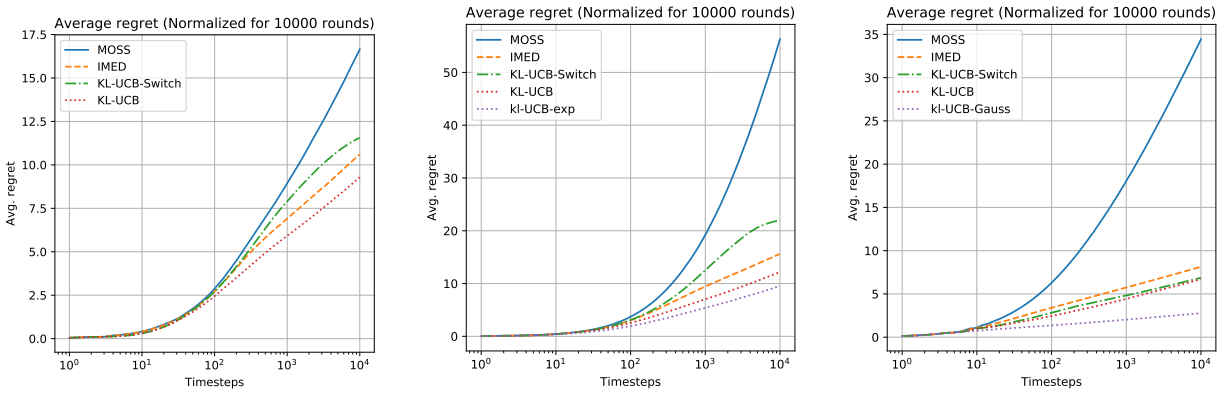

Figure 4.1: Regrets approximated over $10,000$ runs, shown on a log-scale; distributions of the arms consist of:
Left: Bernoulli distributions with parameters (0.9, 0.8)
Middle: Exponential distributions truncated on $[0,1]$, with parameters (0.15, 0.12, 0.1, 0.05)
Right: Gaussian distributions truncated on $[0,1]$, with means (0.7, 0.5, 0.3, 0.2) and same standard deviation $\sigma = 0.1$

Distribution-free bounds. Here we also consider the UCB algorithm of Auer et al. [2002a] with the exploration function $\log(t)$. We plot the behavior of the normalized regret, $R_T/\sqrt{KT}$, either as a function of T (Figure 4.2 left) or of K (Figure 4.2 right). This quantity should not increase without a bounnd as T or K increases. KL-UCB-switch and KL-UCB have a normalized regret that seems to not depend too much on T and K. (KL-UCB may perhaps satisfy a distribution-free bound of the optimal order, but we were unable to prove this fact.) The regret of IMED seems to suffer from a suboptimal dependence in K.

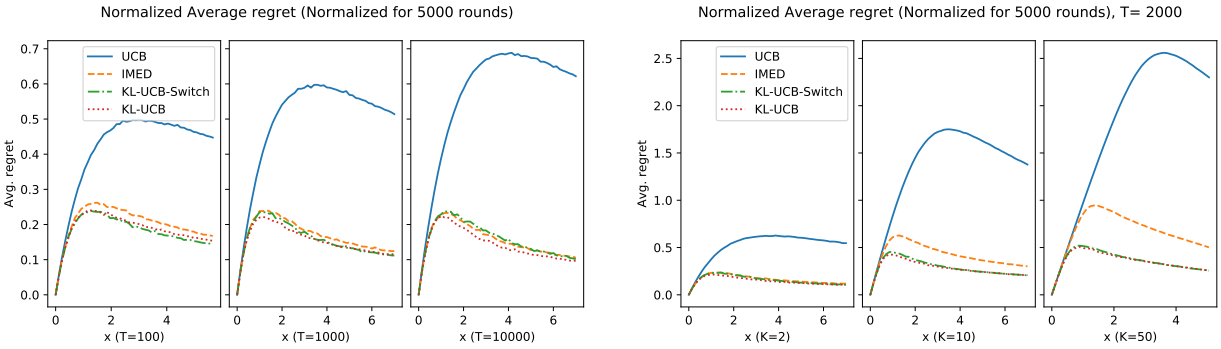

Figure 4.2: Expected regret $R_T/\sqrt{KT}$, approximated over $5,000$ runs
Left: as a function of x, for a Bernoulli bandit problem with parameters $\big(0.8,\, 0.8 - x\sqrt{K/T}\big)$ and for time horizons $T \in \{100,\, 1000,\, 10000\}$
Right: as a function of x, for a Bernoulli bandit problem with parameters $(0.8,\, 0.8 - x\sqrt{K/T},\, \ldots,\, 0.8 - x\sqrt{K/T})$ and K arms, where $K \in \{2,\, 10,\, 50\}$

4.4 Proofs of our main results: the first two theorems of Section 4.2.2

The proof of Theorem 4.2.1 is much standard: it strongly ressembles the proof of MOSS and involves no particular difficulty. Twists had to be considered for the proof of Theorem 4.2.2.

Proof of Theorem 4.2.1. The first step is standard, see Bubeck and Liu [2013]; we use $U_{A_t}(t) \geqslant U_{a^\star}(t)$ to decompose the regret as

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big] + \sum_{t=K+1}^{T} \mathbb{E}\big[U_{A_t}(t) - \mu_{A_t}\big] \quad (4.7)$$

Each term in the second sum in (4.7) is bounded in a crude way: by the application (4.28)

of Pinsker's inequality, $U_a(t) \leqslant U_a^{\mathrm{M}}(t)$ so that

$$\mathbb{E}\big[U_{A_t}(t)-\mu_{A_t}\big] \leqslant \sqrt{\frac{K}{T}}+\mathbb{E}\bigg[\Big(U_{A_t}(t)-\mu_{A_t}-\sqrt{\frac{K}{T}}\Big)^+\bigg] \leqslant \sqrt{\frac{K}{T}}+\mathbb{E}\bigg[\Big(U^{\mathrm{M}}_{A_t}(t)-\mu_{A_t}-\sqrt{\frac{K}{T}}\Big)^+\bigg]$$
$$\leqslant \sqrt{\frac{K}{T}} + \sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\bigg[\Big(U^{\mathrm{M}}_{a,n} - \mu_a - \sqrt{\frac{K}{T}}\Big)^+\bigg]$$

where for the final inequality we used optional skipping (see Section 4.5.1). The first sum in (4.7) is dealt with by substituting the value $U^{\mathrm{KL}}_{a^\star}(t)$ or $U^{\mathrm{M}}_{a^\star}(t)$ of $U_{a^\star}(t)$ depending on $N_{a^\star}(t) \leqslant f(T,K)$ or $N_{a^\star}(t) > f(T,K)$:

$$\sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star}(t)\big] \leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big]$$
$$+ \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t)\big)^+ \underbrace{\mathbb{1}_{\{N_{a^\star}(t)> f(T,K)\}}}_{\leqslant 1}\Big]$$

Collecting the inequalities above into (4.7), we see that the regret of KL-UCB-switch is less than the claimed $(K-1)+25\sqrt{KT}$ bound,

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \underbrace{\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big]}_{\text{we show below that } \leqslant 8\sqrt{K/T} \text{ for each } t}$$
$$\underbrace{+\sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{M}}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 17\sqrt{KT} \quad \text{by (4.25)}}$$

Indeed, by optional skipping (see Section 4.5.1),

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(T,K)\}}\Big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star,n}\big)^+\Big]$$

where by Fubini-Tonelli, for each $1 \leqslant n \leqslant f(T,K) < T/K$,

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{KL}}_{a^\star,n}\big)^+\Big] = \int_0^{\mu^\star} \mathbb{P}\big[\mu^\star - U^{\mathrm{KL}}_{a^\star,n} > u\big]\mathrm{d}u \leqslant \int_0^{+\infty} \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}\mathrm{d}u$$

as for all $u \in (0,\mu^\star)$, by using successively (4.31) and Proposition 4.5.6,

$$\mathbb{P}\big[\mu^\star - U^{\mathrm{KL}}_{a^\star,n} > u\big] \leqslant \mathbb{P}\bigg[\mathcal{K}_{\inf}(\widehat{\nu}_{a^\star,n},\mu^\star) > \frac{1}{n}\log\Big(\frac{T}{Kn}\Big) + 2u^2\bigg] \leqslant \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}$$

The proof of the desired $8\sqrt{K/T}$ bound is concluded by straightforward calculations,

$$\sum_{n=1}^{f(T,K)} \int_0^{+\infty} \mathrm{e}(2n+1)\frac{Kn}{T}\mathrm{e}^{-2nu^2}\mathrm{d}u = \sum_{n=1}^{f(T,K)} \mathrm{e}(2n+1)\frac{Kn}{T}\,\frac{1}{\sqrt{2n}}\sqrt{\frac{\pi}{2}} = \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\sum_{n=1}^{f(T,K)}(2n+1)\sqrt{n}$$
$$\leqslant \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\,f(T,K)\left(2f(T,K)^{3/2}+f(T,K)^{1/2}\right) \leqslant \frac{\mathrm{e}\sqrt{\pi}}{2}\frac{K}{T}\,3f(T,K)^{5/2} \leqslant 8\sqrt{K/T}$$

since by definition of $f(T,K)$, we have $f(T,K)^{5/2} \leqslant (T/K)^{1/2}$. □

The proof of the first distribution-dependent bound (Theorem 4.2.2) relies entirely on elementary applications of concentration inequalities, after some careful cutting of events.

Proof of Theorem 4.2.2. Given $\delta > 0$ sufficiently small (to be determined by the analysis), we decompose $\mathbb{E}\big[N_a(T)\big]$ as

$$\mathbb{E}\big[N_a(T)\big] = 1+\sum_{t=K}^{T-1}\mathbb{P}\big[U_a(t) < \mu^\star-\delta \text{ and } A_{t+1}=a\big]+\sum_{t=K}^{T-1}\mathbb{P}\big[U_a(t) \geqslant \mu^\star-\delta \text{ and } A_{t+1}=a\big] \tag{4.8}$$

Control of the first sum in (4.8). When $A_{t+1}=a$, we have $U_{a^\star}(t) \leqslant U_a(t)$ by definition of the index policy and this is the only piece of information that traditional proofs, as the one of, e.g., Cappé et al. [2013], use: they are left to bound $\sum \mathbb{P}\big[U_{a^\star}(t) < \mu^\star - \delta\big]$. We proceed slightly more carefully by introducing a possibly cutting at $(\mu^\star+\mu_a)/2$ and by distinguishing whether $U_{a^\star}(t)$ is smaller or larger than this value; in the latter case, $U_a(t)$ is also larger than it. In addition we set a threshold $n_0 \geqslant 1$ (to be determined by the analysis) and distinguish whether $N_a(t) \geqslant n_0$ or $N_a(t) \leqslant n_0 - 1$. We thus get the decomposition

$$\begin{aligned}\big\{U_a(t) < \mu^\star - \delta \text{ and } A_{t+1}=a\big\} \subseteq \quad & \big\{U_{a^\star}(t) < (\mu^\star+\mu_a)/2\big\} \\ & \cup \big\{U_a(t) \geqslant (\mu^\star+\mu_a)/2 \text{ and } A_{t+1}=a \text{ and } N_a(t) \geqslant n_0\big\} \\ & \cup \big\{U_{a^\star}(t) < \mu^\star - \delta \text{ and } A_{t+1}=a \text{ and } N_a(t) \leqslant n_0-1\big\}\end{aligned}$$

For $u \in (0,1)$, we introduce the event

$$\mathcal{E}_\star(u) = \Big\{\exists\, \tau \in \{K,\dots,T-1\} : U_{a^\star}(\tau) < u\Big\}$$

We now rewrite the second event in the set decomposition above. To that end, we note that by (4.28) and by definition of the MOSS index, we have, when $N_a(t) \geqslant n_0$,

$$U_a(t) \leqslant U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\log_+\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2n_0}\log_+\left(\frac{T}{Kn_0}\right)}}_{\leqslant \Delta_a/4} \tag{4.9}$$

where the inequality in the root of the right-most term comes from the choice

$$n_0 = \left\lceil \frac{8}{\Delta_a^2} \log\left(\frac{T}{K}\right) \right\rceil \tag{4.10}$$

In particular, we get the inclusion

$$\big\{U_a(t) \geqslant (\mu^\star + \mu_a)/2\big\} = \big\{U_a(t) \geqslant \mu_a + \Delta_a/2\big\} \subseteq \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4\big\}$$

Collecting all elements together and substituting the definition of $\mathcal{E}_\star$, we established the cruder decomposition

$$\begin{aligned}\big\{U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big\} \subseteq \quad & \mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big) \\ & \cup \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4 \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\big\} \\ & \cup \Big(\mathcal{E}_\star(\mu^\star - \delta) \cap \big\{A_{t+1} = a \text{ and } N_a(t) \leqslant n_0 - 1\big\}\Big)\end{aligned}$$

Taking probabilities, resorting to a union bound, summing over t, and using the deterministic control

$$\sum_{t=K}^{T-1} \mathbb{1}_{\{A_{t+1}=a \text{ and } N_a(t) \leqslant n_0 - 1\}} \leqslant n_0$$

we get (and this is where it is so handy that the $\mathcal{E}_\star$ do not depend on a particular t):

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant & \sum_{t=K}^{T-1} \mathbb{P}\left[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\right] \\ & + T\,\mathbb{P}\Big(\mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big)\Big) + n_0\,\mathbb{P}\big(\mathcal{E}_\star(\mu^\star - \delta)\big)\end{aligned} \tag{4.11}$$

By optional skipping (see Section 4.5.1), the first sum above is bounded by

$$\sum_{t=K}^{T-1} \mathbb{P}\left[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\right] \leqslant \sum_{n=n_0}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right]$$

and we continue the upper bounding by applying Hoeffding's inequality (Proposition 4.5.1, actually not using the maximal form):

$$\sum_{n=n_0}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right] \leqslant \sum_{n=n_0}^{T} \mathrm{e}^{-n\Delta_a^2/8} = \frac{\mathrm{e}^{-n_0\Delta_a^2/8}}{1 - \mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K/T}{1 - \mathrm{e}^{-\Delta_a^2/8}} \tag{4.12}$$

where we substituted the value (4.10) of n_0. The two other terms in (4.11) are bounded using the lemma right below, respectively with $x = \Delta_a/2$ and $x = \delta$, and we get the final upper bound

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant & \left(\frac{320\,\mathrm{e}}{(1 - \mathrm{e}^{-2})^3} \frac{K}{\Delta_a^6} + T\mathrm{e}^{-T\Delta_a^2/(2K)}\right) \\ & + \left\lceil \frac{8}{\Delta_a^2} \log\left(\frac{T}{K}\right) \right\rceil \left(\frac{5\mathrm{e}}{(1 - \mathrm{e}^{-2})^3} \frac{K}{T\delta^6} + \mathrm{e}^{-2\delta^2 T/K}\right) + \frac{K/T}{1 - \mathrm{e}^{-\Delta_a^2/8}}\end{aligned} \tag{4.13}$$

The bound above is a $O_T(1)$ for the choices $\delta = (\log T)^{-1/3}$ and $\delta = T^{-1/8}$ respectively considered in Theorems 4.2.2 and 4.2.3.

Lemma 4.4.1. *For all $x \in (0, \mu^\star)$,*

$$\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star - x\big)\Big) = \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : U_{a^\star}(\tau) < \mu^\star - x\Big] \leqslant \frac{5\mathrm{e}}{(1-\mathrm{e}^{-2})^3}\frac{K}{Tx^6} + \mathrm{e}^{-2x^2T/K}$$

Proof. The $\log_+$ in the definition of $U_{a^\star}(\tau)$ vanishes when $N_{a^\star}(\tau) \geqslant T/K$. Therefore, by distinguishing the cases $N_{a^\star}(\tau) < T/K$ and $N_{a^\star}(\tau) \geqslant T/K$, by Pinsker's inequality (4.28), by optional skipping (see Section 4.5.1) and by the definition of the index as a given supremum, we successively get

$$\begin{aligned}
&\mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : U_{a^\star}(\tau) < \mu^\star - x\Big] \\
=&\mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : U_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) < T/K\Big] \\
&+ \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : U_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) \geqslant T/K\Big] \\
=&\mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : U^{\mathrm{KL}}_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) < T/K\Big] \\
&+ \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} : \widehat{\mu}_{a^\star}(\tau) < \mu^\star - x \text{ and } N_{a^\star}(\tau) \geqslant T/K\Big] \\
\leqslant&\mathbb{P}\Big[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} : U^{\mathrm{KL}}_{a^\star,m} < \mu^\star - x\Big] + \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \\
\leqslant&\mathbb{P}\left[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} : \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m}\log\left(\frac{T}{Km}\right)\right] \\
&+ \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big]
\end{aligned}$$

where by a union bound, by the deviation inequality (4.35) stated as a consequence of Proposition 4.5.6, and by some elementary calculations detailed below,

$$\begin{aligned}
&\mathbb{P}\left[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} : \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m}\log\left(\frac{T}{Km}\right)\right] \\
\leqslant \sum_{m=1}^{\lfloor T/K \rfloor} \mathrm{e}(2m+1)\,\frac{Km}{T}\mathrm{e}^{-2mx^2} \leqslant \frac{\mathrm{e}K}{T}\sum_{m=1}^{+\infty} m(2m+1)\,\mathrm{e}^{-2mx^2} \leqslant \frac{5e}{(1-\mathrm{e}^{-2})^3}\frac{K}{Tx^6}
\end{aligned} \tag{4.14}$$

while by Hoeffding's maximal inequality (Proposition 4.5.1)

$$\mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} : \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \leqslant \mathrm{e}^{-2\,\lceil T/K \rceil\, x^2} \leqslant \mathrm{e}^{-2x^2T/K}$$

More precisely, the elementary calculations leading to the final inequality in (4.14) are

based on differentiating the defining series for the exponential distribution: for all $\theta > 0$,

$$\begin{aligned}
\sum_{m=0}^{+\infty} \mathrm{e}^{-m\theta} &= \frac{1}{1-\mathrm{e}^{-\theta}}\,, \\
-\sum_{m=1}^{+\infty} m\,\mathrm{e}^{-m\theta} &= \frac{-\mathrm{e}^{-\theta}}{(1-\mathrm{e}^{-\theta})^2} \geqslant \frac{-1}{(1-\mathrm{e}^{-\theta})^2} \geqslant \frac{-1}{(1-\mathrm{e}^{-\theta})^3}\,, \\
\sum_{m=1}^{+\infty} m^2\,\mathrm{e}^{-m\theta} &= \frac{\mathrm{e}^{-\theta}(1+\mathrm{e}^{-\theta})}{(1-\mathrm{e}^{-\theta})^3} \leqslant \frac{2}{(1-\mathrm{e}^{-\theta})^3}
\end{aligned}$$

Hence

$$\sum_{m=1}^{+\infty} m(2m+1)\,\mathrm{e}^{-m\,2x^2} \leqslant \frac{5}{(1-\mathrm{e}^{-2x^2})^3}$$

Now, since $\theta \in (0,+\infty) \mapsto \theta/(1-\mathrm{e}^{-\theta})$ is increasing and since $2x^2 \leqslant 2$, we have

$$\frac{2x^2}{1-\mathrm{e}^{-2x^2}} \leqslant \frac{2}{1-\mathrm{e}^{-2}} \qquad \text{thus} \qquad \frac{1}{(1-\mathrm{e}^{-2x^2})^3} \leqslant \frac{1}{x^6(1-\mathrm{e}^{-2})^3}$$

which concludes the proof of the final inequality in (4.14), thus the proof of this lemma, and finally, the treatment of the first sum in (4.8). □

Control of the second sum in (4.8). By what are routine manipulations now, namely, distinguishing whether $N_a(t)$ is larger or smaller than $f(T,K)$ and by optional skipping (see Section 4.5.1), we have

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\big[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \delta\big] + \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \tag{4.15}$$

Let us denote

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \log^2\left(\frac{1}{1-\mu^\star}\right)\right)$$

as in the statement of Proposition 4.5.7. For T large enough to satisfy (4.17) and for δ small enough so that $\delta \leqslant \Delta_a/2$ and $\delta^2 \leqslant \gamma_\star(1-\mu^\star)^2/2$, we further upper bound below (4.15) by

$$\frac{K\,f(T,K)/T}{1-\mathrm{e}^{\Delta_a^2/8}} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} + \frac{\log(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \tag{4.16}$$

The obtained bound indeed equals $(\log T)/\mathcal{K}_{\inf}(\nu_a,\mu^\star) + O_T\big((\log T)^{2/3}\big)$ for the considered choice $\delta = (\log T)^{-1/3}$, as can be seen by noting that the first term in (4.16) can be bounded by a constant, while the second and third terms can be dealt with by resorting, respectively, to $1-\mathrm{e}^{-u} = u + o(u)$ and $1/(1-u) = 1 + u + o(u)$ as $u \to 0$, where u is proportional, respectively, to δ^2 and δ.

We turn to the proof of (4.16). We deal with the first sum in the right-hand side of (4.15) as around (4.9): provided that T is large enough, so that

$$\frac{\log\Big(T/\big(K\,f(T,K)\big)\Big)}{f(T,K)} \leqslant \frac{\Delta_a^2}{8}\,, \tag{4.17}$$

we have, for $f(T,K)+1 \leqslant n < T/K$,

$$U_{a,n}^{\mathrm{M}} \leqslant U_{a,\,f(T,K)}^{\mathrm{M}} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2f(T,K)}\log\left(\frac{T}{K\,f(T,K)}\right)} \leqslant \widehat{\mu}_{a,n} + \frac{\Delta_a}{4} \tag{4.18}$$

Note that the $U_{a,n}^{\mathrm{M}} \leqslant \widehat{\mu}_{a,n} + \Delta_a/4$ bound is valid even when $n \geqslant T/K$, as the exploration then vanishes. Therefore, as in (4.12), as soon as $\delta \leqslant \Delta_a/2$,

$$\begin{aligned}
&\sum_{n=f(T,K)+1}^{T} \mathbb{P}\big[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \delta\big] \leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[U_{a,n}^{\mathrm{M}} \geqslant \mu^\star - \frac{\Delta_a}{2}\right] \\
&\leqslant \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} + \frac{\Delta_a}{4} \geqslant \mu^\star - \frac{\Delta_a}{2}\right] = \sum_{n=f(T,K)+1}^{T} \mathbb{P}\left[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\right] \\
&\leqslant \sum_{n=f(T,K)+1}^{T} \mathrm{e}^{-n\Delta_a^2/8} \leqslant \frac{\mathrm{e}^{-f(T,K)\Delta_a^2/8}}{1-\mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K\,f(T,K)/T}{1-\mathrm{e}^{-\Delta_a^2/8}}
\end{aligned} \tag{4.19}$$

where we used again condition (4.17) to get the last inequality.

It only remains to deal with the second sum in the right-hand side of (4.15). Let

$$n_1 = \left\lceil \frac{\log(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \right\rceil$$

For $n_1 \leqslant n \leqslant f(T,K) < T/K$, by definition of the index as a supremum and by left-continuity of $\mathcal{K}_{\inf}$ (see the comments after Lemma 4.5.4),

$$\begin{aligned}
\big\{U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big\} &\subseteq \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\log\left(\frac{T}{Kn}\right)\right\} \\
&\subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)\Big\} \\
&\subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) \leqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star))\Big\}
\end{aligned} \tag{4.20}$$

where the second inclusion only uses $n \geqslant n_1$ and the definition of n_1, and the last inclusion holds by the regularity inequality (4.29). Therefore we may resort to the concentration inequality on $\mathcal{K}_{\inf}$, stated as Proposition 4.5.7: we get, for all $n_1 \leqslant n \leqslant f(T,K)$,

$$\mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \max\left\{\mathrm{e}^{-n/4},\ \exp\left(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\right)\right\}$$

and whether the first or the second argument of the maximum is the largest is independent of n. Therefore, a summation over $n \geqslant n_1$ keeping the latter remark in mind leads to

$$\sum_{n=n_1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \max\left\{ \frac{1}{1-\mathrm{e}^{-1/4}}, \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} \right\} \tag{4.21}$$

Now if $\delta^2 \leqslant \gamma_\star(1-\mu^\star)^2/2$ we may keep only the second term in the maximum. For such δ, by bounding by 1 the first $n_1 - 1$ probabilities in the sum in (4.15), we finally obtain

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{\log(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} \tag{4.22}$$

which concludes the proof of (4.16). □

4.5 Results (almost) extracted from the literature

We gather in this section results that are all known and published elsewhere (or almost). For the sake of self-completeness we provide a proof of each of them (sometimes this proof is shorter or simpler than the known proofs, and we then comment on this fact).

4.5.1 Optional skipping

The trick detailed here is much standard in the bandit literature, see, e.g., its application in Auer et al. [2002a].

We detail how to reindex various quantities like $U_a(t)$, $\widehat{\mu}_a(t)$, etc., that are indexed by the global time t, into versions indexed by the local number of times $N_a(t) = n$ the specific arm considered has been pulled. The corresponding quantities will be denoted by $U_{a,n}$, $\widehat{\mu}_{a,n}$, etc.

The reindexation is possible as soon as the considered algorithm pulls each arm infinitely often; it is the case for all algorithms considered in this chapter (exploration never stops even if it becomes rare after a certain time).

We denote by $\mathcal{F}_0 = \{\emptyset, \Omega\}$ the trivial σ–algebra and by $\mathcal{F}_t$ the σ–algebra generated by $A_1, Y_1, \ldots, A_t, Y_t$, when $t \geqslant 1$. We fix an arm a. For each $n \geqslant 1$, we denote by

$$\tau_{a,n} = \min\big\{t \geqslant 1 : \quad N_a(t) = n\big\}$$

the round at which arm a was pulled for the n–th time. Doob's optional skipping (see, e.g., Chow and Teicher, 1988, Section 5.3 for a reference) ensures that the random variables $X_{a,n} = Y_{\tau_{a,n}}$ are independent and identically distributed according to ν_a.

We can then define, for instance, for $n \geqslant 1$,

$$\widehat{\mu}_{a,n} = \frac{1}{n}\sum_{k=1}^{n} X_{a,k}$$

and have the equality $\widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)}$ for $t \geqslant K$. Here is an example of how to use this rewriting. Recall that $N_a(t) \geqslant 1$ for $t \geqslant K$ and $N_a(t) \geqslant t$, even $N_a(t) \geqslant t-K+1$ as each arm was pulled once in the first rounds. Given a subset $\mathcal{E} \subseteq [0,1]$, we get the inclusion

$$\big\{\widehat{\mu}_a(t) \in \mathcal{E}\big\} = \bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_a(t) \in \mathcal{E} \text{ and } N_a(t) = n\big\} = \bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_{a,n} \in \mathcal{E} \text{ and } N_a(t) = n\big\}$$

so that, by a union bound,

$$\mathbb{P}\big[\widehat{\mu}_a(t) \in \mathcal{E}\big] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E} \text{ and } N_a(t) = n\big] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E}\big]\,.$$

The last sum above only deals with independent and identically distributed random variables; we took care of all dependency issues that are so present in bandit problems. The price to pay, however, is that we bounded one probability by a sum of probabilities.

4.5.2 Maximal Hoeffding's inequality

This much standard result from Hoeffding [1963] was already used in the proof of the regret bound of MOSS (Audibert and Bubeck, 2009).

Proposition 4.5.1. *Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables bounded in $[0,1]$ and let $\widehat{\mu}_n$ denote their empirical mean. Then for all $u > 0$ and for all $N \geqslant 1$:*

$$\mathbb{P}\Big[\max_{n \geqslant N} \big(\widehat{\mu}_n - \mu\big) \geqslant u\Big] \leqslant \mathrm{e}^{-2Nu^2} \tag{4.23}$$

Corollary 4.5.2. *Under the same assumptions, for all $\varepsilon > 0$,*

$$\mathbb{E}\left[\Big(\max_{n \geqslant N} \big(\mu - \widehat{\mu}_n - \varepsilon\big)\Big)^+\right] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2} \tag{4.24}$$

Proof. By Fubini-Tonelli, an integration of the maximal concentration inequality yields

$$\begin{aligned}\mathbb{E}\left[\Big(\max_{n \geqslant N} \big(\mu - \widehat{\mu}_n - \varepsilon\big)\Big)^+\right] &= \int_0^{+\infty} \mathbb{P}\Big[\max_{n \geqslant N} \big(\widehat{\mu}_n - \mu - \varepsilon\big) \geqslant u\Big]\mathrm{d}u \\ &\leqslant \int_0^{+\infty} \mathrm{e}^{-2N(u+\varepsilon)^2}\mathrm{d}u \leqslant \mathrm{e}^{-2N\varepsilon^2}\int_0^{+\infty} \mathrm{e}^{-2Nu^2}\mathrm{d}u = \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2}\end{aligned}$$

□

4.5.3 Analysis of the MOSS algorithm

This analysis was already performed in the literature, both for a known horizon T (see Audibert and Bubeck 2009) and for an anytime version (see Degenne and Perchet 2016). We provide slightly shorter and more focused proofs of these results based on Proposition 4.5.2 in Appendix 4.7.2; the main difference to the mentioned proofs lies

in elegance. Typically, the peeling trick was used on the probabilities of deviations (see Proposition 4.5.1) and had to be performed separately and differently for each deviation u; then, these probabilities were integrated to obtain a control on the needed expectations. In contrast, we perform the peeling trick directly on the expectations at hand, and we do so by applying it only once, at fixed times depending solely on T, which makes the proof more readable. Put differently, we do not claim any improvement on the results themselves, just a clarification of their proof. This proof is also very similar to the one of the minimax optimality of Theorem 3.3.5 in the parametric setting of Chapter 3.

We first recall the distribution-free bound on the regret of MOSS, when T is known. We also extract an intermediary result from its proof, which will be used in the analysis of our algorithm. We denote by A_t^{M} the arm played by the index strategy maximizing, at each step $t+1$ with $t \geqslant K$, the quantity

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\left(\frac{T}{KN_a(t)}\right)}$$

Proposition 4.5.3. *For all bandit problems $\underline{\nu}$ over $[0,1]$, the regret of MOSS satisfies*

$$R_T \leqslant (K-1) + 17\sqrt{KT}$$

More precisely, we have the inequalities

$$R_T - (K-1) \leqslant \sqrt{KT} + \underbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\mathrm{M}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n}^{\mathrm{M}} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 16\sqrt{KT}} \tag{4.25}$$

Our proof in Appendix 4.7.2 reveals that designing an adaptive version of MOSS comes at no effort; indeed, MOSS-anytime relies on the indexes, for $t \geqslant K$,

$$U_a^{\text{M-A}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\left(\frac{t}{KN_a(t)}\right)} \tag{4.26}$$

and satisfies a regret bound of $(K-1) + 29\sqrt{KT}$.

4.5.4 Regularity and deviation/concentration results on $\mathcal{K}_{\inf}$

Many results of this section rely on Pinsker's inequality. One of its most basic consequences is in terms of a lower bound on $\mathcal{K}_{\inf}$. Indeed, since we are considering distributions over $[0,1]$, the data-processing inequality for Kullback-Leibler divergences ensures (see, e.g., Lemma 2.2.3 that for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in \big(\mathrm{E}(\nu), 1\big)$,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \inf_{\nu' : \mathrm{E}(\nu') > \mu} \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}\big(\mathrm{E}(\nu')\big)\Big) = \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mu\Big),$$

where $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. Therefore, by Pinsker's inequality for Bernoulli distributions,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant 2\big(\mathrm{E}(\nu)-\mu\big)^2\,, \qquad \text{thus} \qquad U_a^{\mathrm{KL}}(t) \leqslant U_a^{\mathrm{M}}(t) \tag{4.27}$$

for all arms a and all rounds $t \geqslant K$. In particular, for KL-UCB-switch,

$$U_a^{\mathrm{KL}}(t) \leqslant U_a(t) \leqslant U_a^{\mathrm{M}}(t) \tag{4.28}$$

Another consequence of Pinsker's inequality is given by the inequality (4.30) below, while the inequality (4.29) appears as Lemma 7 in Honda and Takemura [2015]. These two inequalities are proved in details in Section 4.7.4; the proposed proofs are slightly simpler or lead to sharper bounds than in the mentioned references.

Lemma 4.5.4 (regularity of $\mathcal{K}_{\inf}$)**.** *For all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in (0,1)$,*

$$\forall \varepsilon \in (0,\mu)\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + \frac{\varepsilon}{1-\mu}\,, \tag{4.29}$$

and

$$\forall \varepsilon \in \big[0,\mu-\mathrm{E}(\nu)\big]\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon^2\,. \tag{4.30}$$

A consequence of (4.5.4) is the left-continuity of $\mathcal{K}_{\inf}$: for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in (0,1)$, we have $\mathcal{K}_{\inf}(\nu,\mu-\varepsilon) \nearrow \mathcal{K}_{\inf}(\nu,\mu)$ as $\varepsilon \searrow 0$. Therefore, by a sandwich argument, $\mathcal{K}_{\inf}\big(\nu,\mathrm{E}(\nu)\big) = 0$ whenever $\mathrm{E}(\nu) \in (0,1)$.

A consequence of (4.30) is the following. For all $B > 0$, all $\widetilde{\mu} \in (0,1)$, all $\varepsilon \in [0,\widetilde{\mu})$, and all distributions ν over $[0,1]$ with $\mathrm{E}(\nu) < \widetilde{\mu}-\varepsilon$,

$$\Big\{\sup\big\{\mu \in [0,1] \;\big|\; \mathcal{K}_{\inf}(\nu,\mu) \leqslant B\big\} < \widetilde{\mu}{-}\varepsilon\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\nu,\widetilde{\mu}{-}\varepsilon\big) > B\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\nu,\widetilde{\mu}\big) > B{+}2\varepsilon^2\Big\}\,, \tag{4.31}$$

and these inclusions still hold even when $\mathrm{E}(\nu) \geqslant \widetilde{\mu}-\varepsilon$, as in this case, the left-most set is empty.

The variational formula appears in Honda and Takemura [2015] as Theorem 2 (and Lemma 6) and is an essential tool for deriving concentration results for the $\mathcal{K}_{\inf}$. We re-derive it in an elegant and direct way in Section 4.7.4.

Lemma 4.5.5 (variational formula for $\mathcal{K}_{\inf}$)**.** *For all $\nu \in \mathcal{P}[0,1]$ and all $0 < \mu < 1$,*

$$\mathcal{K}_{\inf}(\nu,\mu) = \max_{0\leqslant\lambda\leqslant 1} \mathbb{E}\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right] \qquad \textit{where } X \sim \nu \tag{4.32}$$

Moreover, if we denote by $\lambda^\star$ the value at which the above maximum is reached, then

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star(X-\mu)/(1-\mu)}\right] \leqslant 1 \tag{4.33}$$

The following deviation inequality on $\mathcal{K}_{\inf}$ was provided by Cappé et al. [2013, Lemma 6] in all cases where the variational formula (4.32) holds. For the sake of completeness, we recall its proof in Section 4.7.4.

Proposition 4.5.6 (deviation result on $\mathcal{K}_{\inf}$)**.** *Let $\widehat{\nu}_n$ denote the empirical distribution associated with a sequence of n i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geqslant 0$,*

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-nu} \tag{4.34}$$

A consequence of the proposition above and of Lemma 4.5.4 is the following one: for all $u > 0$ and all $\varepsilon \in \big[0, \mathrm{E}(\nu)\big)$,

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) \geqslant u\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n(u+2\varepsilon^2)} \tag{4.35}$$

Indeed, when ε is such that $\mathrm{E}(\nu) - \varepsilon < \widehat{\mu}_n$, where $\widehat{\mu}_n$ denotes the average of the considered i.i.d. random variables, then $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) = 0$ by definition, while otherwise, by (4.30), since $\varepsilon \in \big[0, \mathrm{E}(\nu) - \widehat{\mu}_n\big]$, we have

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) + 2\varepsilon^2 \leqslant \mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)\,.$$

Therefore, the inclusion

$$\Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu) - \varepsilon\big) \geqslant u\Big\} \subseteq \Big\{\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u + 2\varepsilon^2\Big\}$$

is valid for all $u > 0$ and (4.35) follows from Proposition 4.5.6.

The next proposition is similar in spirit to Honda and Takemura [2015, Proposition 11] but is better suited to our needs. We prove it in Appendix 4.7.1.

Proposition 4.5.7 (concentration result on $\mathcal{K}_{\inf}$)**.** *With the same notation and assumptions as in the previous proposition, consider a real number $\mu^\star \in \big(\mathrm{E}(\nu), 1\big)$ and define*

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \log^2\left(\frac{1}{1-\mu^\star}\right)\right) \tag{4.36}$$

Then for all $x < \mathcal{K}_{\inf}(\nu, \mu^\star)$,

$$\mathbb{P}\big[\mathcal{K}_{\inf}(\widehat{\nu}_n, \mu^\star) \leqslant x\big] \leqslant \begin{cases} \exp(-n\gamma_\star/8) \leqslant \exp(-n/4) & \text{if } x \leqslant \mathcal{K}_{\inf}(\nu, \mu^\star) - \gamma_\star/2 \\ \exp\Big(-n\big(\mathcal{K}_{\inf}(\nu, \mu^\star) - x\big)^2/(2\gamma_\star)\Big) & \text{if } x > \mathcal{K}_{\inf}(\nu, \mu^\star) - \gamma_\star/2 \end{cases}$$

4.6 Proof of the more advanced bound of Theorem 4.2.3

The proof of the sharper bound of Theorem 4.2.3 relies on the following lemma, which was (almost) stated in Honda and Takemura [2015, Lemma 18]: our assumptions and result are slightly different (they are tailored to our needs), which is why we provide below a proof of this lemma.

By convention, the infimum over an empty set equals $+\infty$. In what follows, $\wedge$ denotes the minimum of two numbers; the considered stopping time τ is thus always bounded by T. We recall that Lambert's function W is defined, for $x > 0$, as the unique solution $W(x)$ of the equation $w\,\mathrm{e}^w = x$, with unknown $w > 0$. We recall (see, e.g., Hoorfar and Hassani, 2008a, Corollary 2.4) that it is increasing and that

$$\forall\, x > \mathrm{e}, \qquad \log x - \log\log x \leqslant W(x) \leqslant \log x - \log\log x + \log\left(1 + \mathrm{e}^{-1}\right) \tag{4.37}$$

and in particular, $W(x) = \log x - \log\log x + O(1)$ as $x \to +\infty$.

Lemma 4.6.1. *Let (Z_i) be a sequence of i.i.d. variables with a positive expectation $\mathbb{E}[Z_1] > 0$ and such that $Z_i \leqslant \alpha$ for some $\alpha > 0$. For an integer $T \geqslant 1$, consider the stopping time*

$$\tau \stackrel{\text{def}}{=} \inf\left\{ n \geqslant 1 \;\middle|\; \sum_{i=1}^{n} Z_i > \log\left(\frac{T}{Kn}\right) \right\} \wedge T$$

Then, for all $T \geqslant K\mathrm{e}^{\alpha}$,

$$\mathbb{E}[\tau] \leqslant \frac{W(\alpha T/K) + \alpha + \log 2}{\mathbb{E}[Z_1]}$$

where W is Lambert's function.

Proof. We consider the martingale $(M_n)_{n\geqslant 0}$ defined by

$$M_n = \sum_{i=1}^{n} \big(Z_i - \mathbb{E}[Z_1]\big)$$

As τ is a finite stopping time, Doob's optional stopping theorem indicates that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0] = 0$, that is,

$$\mathbb{E}[\tau]\,\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{i=1}^{\tau} Z_i\right]$$

That first step of the proof was similar to the one of Honda and Takemura [2015, Lemma 18]. The idea is now to upper bound the right-hand side of the above equality, which we do by resorting to the very definition of τ. An adaptation is needed with respect to the original argument as the value $\log\big(T/(Kn)\big)$ of the barrier varies with n.

We proceed as follows. Since $Z_1 \leqslant \alpha$ and $T \geqslant K\mathrm{e}^{\alpha}$ by assumption, we necessarily have $\tau \geqslant 2$; using again the boundedness by α, we have, by definition of τ,

$$\sum_{i=1}^{\tau-1} Z_i + Z_\tau \leqslant \log\left(\frac{T}{K(\tau-1)}\right) + \alpha = \log\left(\frac{T}{K\tau}\right) + \log\left(\frac{\tau}{\tau-1}\right) + \alpha \leqslant \log\left(\frac{T}{K\tau}\right) + \log 2 + \alpha$$

In addition, when $\tau < T/K$,

$$\log\left(\frac{T}{K\tau}\right) < \sum_{i=1}^{\tau} Z_i \leqslant \tau\alpha \qquad \text{thus} \qquad 0 < \frac{T}{K\tau}\log\left(\frac{T}{K\tau}\right) \leqslant \frac{T\alpha}{K}$$

Applying the increasing function W to all sides of the latter inequality, we get, when $\tau < T/K$,

$$\log\left(\frac{T}{K\tau}\right) \leqslant W\left(\frac{T\alpha}{K}\right)$$

This inequality also holds when $\tau \geqslant T/K$ as the left-hand side then is non-positive, while the right-hand side is positive. Putting all elements together, we successively proved

$$\mathbb{E}[\tau]\,\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{i=1}^{\tau} Z_i\right] \leqslant \mathbb{E}\left[\log\left(\frac{T}{K\tau}\right)\right] + \log 2 + \alpha \leqslant W\left(\frac{T\alpha}{K}\right) + \log 2 + \alpha$$

which concludes the proof. □

Proof of Theorem 4.2.3. All inequalities of the proof of Theorem 4.2.2 hold in the present case as well, given that we are studying exactly the same algorithm. The regret is decomposed as in the mentioned proof, and inequality (4.13) holds as a first part of the final regret bound. Now, the second part consists of (4.15), which we bound as (4.19) plus the bound

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\log\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \log\big(2/(1-\mu^\star)\big)\right) + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \quad (4.38)$$

where again,

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}}\left(16\mathrm{e}^{-2} + \log^2\left(\frac{1}{1-\mu^\star}\right)\right)$$

To do so, we use the conditions (4.17) and $T > K/(1-\mu^\star)$ on T, and the conditions $\delta \leqslant \Delta_a/2$ and $\delta \leqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star)/\big(2(1-\mu^\star)\big)$ on δ; all in all, we get

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a\big] \leqslant \frac{K\,f(T,K)/T}{1-\mathrm{e}^{\Delta_a^2/8}} + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} + \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\log\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \log\big(2/(1-\mu^\star)\big)\right) \quad (4.39)$$

Since $1/(1-u) = 1 + O(u)$ as $u \to 0$, for the choice $\delta = T^{-1/8}$ contemplated in Theorem 4.2.3, the bound above equals

$$\frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)} W\left(\frac{\log\big(1/(1-\mu^\star)\big)}{K}\, T\right) + O_T(1)$$
$$= \frac{W(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\big(1 + O_T(T^{-1/8})\big) + O_T(1) = \frac{\log T - \log\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + O_T(1)\,,$$

where the final equality follows from the asymptotic expansion (4.37).

The difference with the proof of Theorem 4.2.2 lies in a sharper bound of the quantity (4.38), given by the last two terms in the above inequality (4.39). We follow exactly the same method as in the analysis of the IMED policy of Honda and Takemura [2015, Theorem 5]: their idea was to deal with the deviations in a more careful way and relate the sum (4.38) to the behaviour of a biased random walk.

We start by following the same steps as in the proof of Proposition 4.5.7 in Appendix 4.7.1 and link the deviations in $\mathcal{K}_{\inf}$ divergence to the ones of a random walk. The variational formulation (Lemma 4.5.5) for $\mathcal{K}_{\inf}$ entails the existence of $\lambda_{a,\delta} \in [0,1]$ such that

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) = \mathbb{E}\left[\log\left(1 - \lambda_{a,\delta}\frac{X_a - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right)\right] \qquad \text{where} \qquad X_a \sim \nu_a$$

Note that $\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) > 0$ by (4.27) given that $\delta \leqslant \Delta_a/2$. We consider i.i.d. copies $X_{a,1}, \dots, X_{a,n}$ of X and form the random variables

$$Z_{a,i} = \log\left(1 - \lambda_{a,\delta}\frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right)$$

where, since $X_{a,i} \geqslant 0$ and $\lambda_{a,\delta} \in [0,1]$, we have

$$Z_{a,i} = \log\left(1 - \lambda_{a,\delta}\frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right) \leqslant \log\left(1 + \lambda_{a,\delta}\frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right)$$
$$\leqslant \log\left(1 + \frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right) = \log\left(\frac{1}{1 - (\mu^\star - \delta)}\right) \leqslant \log\left(\frac{1}{1-\mu^\star}\right) \stackrel{\text{def}}{=} \alpha$$

By the variational formulation again, applied this time to $\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta)$,

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \geqslant \frac{1}{n}\sum_{i=1}^{n} Z_{a,i}$$

which entails, for each $n \geqslant 1$,

$$\big\{U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big\} \subseteq \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\log\left(\frac{T}{Kn}\right)\right\} \subseteq \left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log\left(\frac{T}{Kn}\right)\right\} \tag{4.40}$$

where the first inclusion holds for the same reasons (including left-continuity of $\mathcal{K}_{\mathrm{inf}}$) as in (4.20). Therefore, the quantity of interest (4.38) is bounded by

$$\begin{aligned}\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] &\leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\sum_{i=1}^{n} Z_{a,i} \leqslant \log\left(\frac{T}{Kn}\right)\right] \\ &= \mathbb{E}\left[\sum_{n=1}^{f(T,K)} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log(T/(Kn))\right\}}\right] \\ &\leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log(T/(Kn))\right\}}\right]\end{aligned}$$

This latter sum can be reinterpreted as the expected number of times a random walk with positive bias stays under a decreasing logarithmic barrier. We exploit this interpretation to our advantage by decomposing this sum into the expected hitting time of the barrier and a sum of deviation probabilities for the walk.

Let us therefore define the first hitting time τ of the barrier

$$\tau = \inf\left\{n \geqslant 1 \;\middle|\; \sum_{i=1}^{n} Z_{a,i} > \log\left(\frac{T}{Kn}\right)\right\} \wedge T \tag{4.41}$$

which is is a stopping time with respect to the filtration generated by the family $(Z_{a,i})_{1\leqslant i\leqslant n}$. By distinguishing according to whether or not the condition in the defining infimum of τ is met for a $1 \leqslant n \leqslant T$ or not, i.e., whether or not the barrier is hit for $1 \leqslant n \leqslant T$, we get

$$\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log(T/(Kn))\right\}}\right] \leqslant \mathbb{E}[\tau] + \mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log(T/(Kn))\right\}}\right] \tag{4.42}$$

where the sum from $\tau+1$ to T is void thus null when $\tau = T$ (this is the case, in particular, when the barrier is hit for no $n \leqslant T$). Lemma 4.6.1 applies, as, among others, $Z_{a,i} \leqslant \alpha = \log\big(1/(1-\mu^\star)\big)$ as shown above and $T > K/(1-\mu^\star)$; it yields

$$\mathbb{E}[\tau] \leqslant \frac{1}{\mathcal{K}_{\mathrm{inf}}(\nu_a, \mu^\star - \delta)}\left(W\left(\frac{\log\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \log\big(2/(1-\mu^\star)\big)\right)$$

We apply the regularity inequality on $\mathcal{K}_{\mathrm{inf}}$, see also (4.45) below, to get the claimed bound on the first part of (4.42) We now bound its second part. We may assume that $\tau < T$ so that

$$\log\left(\frac{T}{K\tau}\right) < \sum_{i=1}^{\tau} Z_{a,i}$$

For $n \geqslant \tau$, we then have

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \log\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=1}^{n} Z_{a,i} \leqslant \log\left(\frac{T}{K\tau}\right) \leqslant \sum_{i=1}^{\tau} Z_{a,i} \tag{4.43}$$

Hence, in this case,

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \log\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=\tau+1}^{n} Z_{a,i} \leqslant 0\,.$$

This, together with a breakdown according to the values of τ and the independence between $\{\tau = k\}$ and $X_{k+1}, \dots, X_T$, yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \log(T/(Kn))\right\}}\right] \\
&\leqslant \mathbb{E}\left[\sum_{n=\tau+1}^{T} \mathbb{1}_{\left\{\sum_{i=\tau+1}^{n} Z_{a,i} \leqslant 0\right\}}\right] = \sum_{k=1}^{T} \mathbb{E}\left[\mathbb{1}_{\{\tau=k\}} \sum_{n=k+1}^{T} \mathbb{1}_{\left\{\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right\}}\right] \\
&= \sum_{k=1}^{T} \sum_{n=k+1}^{T} \mathbb{P}[\tau = k]\, \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] \\
&= \sum_{k=1}^{T} \mathbb{P}[\tau = k] \left(\underbrace{\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right]}_{\text{we show below } \leqslant \beta}\right) \leqslant \beta \overset{\text{def}}{=} 5 + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \qquad (4.44)
\end{aligned}$$

Indeed, by the concentration results on $\mathcal{K}_{\inf}$ (Proposition 4.5.7), denoting

$$\gamma_{\star,\delta} = \frac{1}{\sqrt{1 - (\mu^\star - \delta)}}\left(16\mathrm{e}^{-2} + \log^2\left(\frac{1}{1 - (\mu^\star - \delta)}\right)\right) \leqslant \gamma_\star\,,$$

we get

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] &\leqslant \max\left\{\mathrm{e}^{-(n-k)/4},\ \exp\left(-\frac{n-k}{2\gamma_{\star,\delta}}\Big(\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)\Big)^2\right)\right\} \\
&\leqslant \mathrm{e}^{-(n-k)/4} + \exp\left(-\frac{n-k}{2\gamma_\star}\Big(\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)\Big)^2\right) \\
&\leqslant \mathrm{e}^{-(n-k)/4} + \mathrm{e}^{-(n-k)\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}
\end{aligned}$$

where the third inequality follows from the first regularity inequality of Lemma 4.5.4 and from our stated condition $\delta \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star)/\big(2(1-\mu^\star)\big)$:

$$\mathcal{K}_{\inf}(\nu, \mu^\star - \delta) \geqslant \mathcal{K}_{\inf}(\nu, \mu^\star) - \frac{\delta}{1 - \mu^\star} \geqslant \frac{\mathcal{K}_{\inf}(\nu, \mu^\star)}{2} \qquad (4.45)$$

We finally get, after summation over $n = k+1, \dots, T$,

$$\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\right] \leqslant \underbrace{\frac{1}{1 - \mathrm{e}^{-1/4}}}_{\leqslant 5} + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}}\,,$$

which is the inequality claimed in (4.44). □

4.7 Elements of Proof

4.7.1 Proof of Proposition 4.5.7

The proof of Proposition 4.5.7 relies on the following lemma via the variational formula (4.32). This lemma is a concentration result for random variables that are essentially bounded from one side only. Its holds also for possibly negative u (there is no lower bound on the u that can be considered).

Lemma 4.7.1. *Let $Z_1, \dots, Z_n$ be i.i.d. random variables such that there exist a, $b \geqslant 0$ with*

$$Z_1 \leqslant a \quad a.s. \quad \text{and} \quad \mathbb{E}\big[\mathrm{e}^{-Z_1}\big] \leqslant b$$

Define furthermore $\gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$. Then Z_1 in integrable and for all $u < \mathbb{E}[Z_1]$,

$$\mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nu\right] \leqslant \begin{cases} \exp(-n\gamma/8) & \text{if } u \leqslant \mathbb{E}[Z_1] - \gamma/2 \\ \exp\Big(-n\big(\mathbb{E}[Z_1] - u\big)^2/(2\gamma)\Big) & \text{if } u > \mathbb{E}[Z_1] - \gamma/2 \end{cases}$$

Indeed, denoting by $\lambda^\star \in [0,1]$ a real number achieving the maximum in the variational formula (4.32) for $\mathcal{K}_{\inf}(\nu, \mu^\star)$, we introduce the random variable

$$Z = \log\left(1 - \lambda^\star \frac{X - \mu^\star}{1 - \mu^\star}\right) \qquad \text{where} \qquad X \sim \nu$$

and i.i.d. copies $Z_1, \dots, Z_n$ of Z. Then, $\mathcal{K}_{\inf}(\nu, \mu^\star) = \mathbb{E}\big[Z\big]$ and by the variational formula (4.32) again,

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mu^\star\big) \geqslant \frac{1}{n}\sum_{i=1}^n Z_i\,, \qquad \text{therefore,} \qquad \mathbb{P}\big[\mathcal{K}_{\inf}(\widehat{\nu}_n, \mu^\star) \leqslant x\big] \leqslant \mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nx\right]$$

for all real numbers x. Now,

$$X \geqslant 0 \qquad \text{thus} \qquad Z \leqslant \log\left(1 + \lambda^\star \frac{\mu^\star}{1 - \mu^\star}\right) \leqslant \log\left(\frac{1}{1 - \mu^\star}\right) \stackrel{\text{def}}{=} a$$

and on the other hand,

$$\mathbb{E}\big[\mathrm{e}^{-Z}\big] = \mathbb{E}\left[\frac{1}{1 - \lambda^\star(X - \mu^\star)/(1 - \mu^\star)}\right] \stackrel{\text{def}}{=} b$$

where $b \leqslant 1$ follows from (4.33). This proves Lemma 4.7.1, except for the inequality $\mathrm{e}^{-n\gamma_\star/8} \leqslant \mathrm{e}^{-n/4}$ claimed therein. The latter is a consequence of $\gamma_\star \geqslant 2$, as $\gamma_\star$ is an increasing function of $\mu^\star > 0$,

$$\gamma_\star = \frac{1}{\sqrt{1 - \mu^\star}}\left(16\mathrm{e}^{-2} + \log^2\left(\frac{1}{1 - \mu^\star}\right)\right) > 16\mathrm{e}^{-2} > 2\,.$$

Proof of Lemma 4.7.1

For the sake of completeness, we provide a proof of Lemma 4.7.1 which is a direct application of the Crámer-Chernoff method.

Proof. We will make repeated uses of the fact that e^{-Z_1} is integrable (by the assumption on b), and that so is e^{Z_1}, as e^{Z_1} takes bounded values in $(0, \mathrm{e}^a]$. In particular, Z_1 is integrable, as by Jensen's inequality,

$$\mathbb{E}\big[|Z_1|\big] \leqslant \log \mathbb{E}\Big[\mathrm{e}^{|Z_1|}\Big] \leqslant \log\Big(\mathbb{E}\big[\mathrm{e}^{-Z_1}\big] + \mathbb{E}\big[\mathrm{e}^{Z_1}\big]\Big) < +\infty$$

We will show below that the log-moment generation function Λ of Z_1 is well-defined at least on the interval $[-1, 1]$,

$$\Lambda : x \in [-1, 1] \longmapsto \log \mathbb{E}\big[\mathrm{e}^{xZ_1}\big]$$

and twice differentiable at least on $(-1, 1)$, with $\Lambda'(0) = \mathbb{E}[Z_1]$ and $\Lambda''(x) \leqslant \gamma$ for $x \in [-1/2,\, 0]$. By a Taylor expansion with a Cauchy remainder, we then have

$$\forall\, x \in [-1/2,\, 0], \qquad \Lambda(x) \leqslant \Lambda(0) + x\,\Lambda'(0) + \frac{x^2}{2} \sup_{y \in (-1/2,\, 0)} \Lambda''(y) \leqslant x\,\mathbb{E}[Z_1] + \frac{\gamma}{2} x^2$$

Therefore, by the Crámer-Chernoff method, for all $x \in [-1/2,\, 0]$, the probability of interest is bounded by

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nu\right] = \mathbb{P}\left[\prod_{i=1}^n \mathrm{e}^{xZ_i} \geqslant \mathrm{e}^{nux}\right] &\leqslant \mathrm{e}^{-nux} \Big(\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]\Big)^n = \exp\Big(-n\big(ux - \Lambda(x)\big)\Big) \\ &\leqslant \exp\left(-n \min_{x \in [-1/2, 0]} \Big\{x\,\big(u - \mathbb{E}[Z_1]\big) - x^2\,\gamma/2\Big\}\right) \end{aligned} \tag{4.46}$$

which we will further upper bound depending on whether $u > \mathbb{E}[Z_1] - \gamma/2$ or $u \leqslant \mathbb{E}[Z_1] - \gamma/2$.

Proofs of the statements on Λ. That Λ is well-defined over $[-1, 1]$ follows from the inequality $\mathrm{e}^{xZ_1} \leqslant \mathrm{e}^{Z_1} + \mathrm{e}^{-Z_1}$, which is valid for all $x \in [-1, 1]$ and whose right-hand side is integrable as already noted above. That $\psi : x \mapsto \mathbb{E}\big[\mathrm{e}^{xZ_1}\big]$ is differentiable at least on $(-1, 1)$ follows from the fact that $x \in (-1, 1) \mapsto Z_1\,\mathrm{e}^{xZ_1}$ is locally dominated by an integrable random variable; indeed, for $x \in (-1, 1)$,

$$\big|Z_1\,\mathrm{e}^{xZ_1}\big| = Z_1\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1 \geqslant 0\}} + Z_1\,\mathrm{e}^{xZ_1}\,\mathbb{1}_{\{Z_1 < 0\}} \leqslant a\,\mathrm{e}^a + \frac{1}{x} \sup_{(-\infty, 0)} f = a\,\mathrm{e}^a + \frac{1}{\mathrm{e}x}$$

where $f(t) = -t\,\mathrm{e}^t$. Similarly, $x \in (-1, 1) \mapsto Z_1^2\,\mathrm{e}^{xZ_1}$ is also locally dominated by an integrable random variable. Thus, ψ is twice differentiable at least on $(-1, 1)$, with first and second derivatives

$$\psi'(x) = \mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big] \qquad \text{and} \qquad \psi''(x) = \mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]$$

and therefore, so is $\Lambda = \log \psi$, with

$$\Lambda'(x) = \frac{\psi'(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]} \quad \text{and} \quad \Lambda''(x) = \frac{\psi''(x)\,\psi(x) - \big(\psi'(x)\big)^2}{\psi(x)^2} \leqslant \frac{\psi''(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]}$$

In particular, $\Lambda'(0) = \mathbb{E}[Z_1]$. As for the bound on $\Lambda''(x)$, we note first that $\mathrm{e}^{xZ_1} \geqslant \mathrm{e}^{xa} \geqslant 1/\sqrt{\mathrm{e}^a}$ as $Z_1 \leqslant a$ and $x \in [-1/2,\, 0]$. Second, using that (proof below)

$$\forall\, x \in [-1/2,\, 0],\ z \in (-\infty, a), \qquad z^2\,\mathrm{e}^{xz} \leqslant 16\,\mathrm{e}^{-2}\mathrm{e}^{-z} + a^2 \tag{4.47}$$

we get $\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big] \leqslant 16\,\mathrm{e}^{-2}b + a^2$. The claimed bound $\Lambda''(x) \leqslant \gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$ follows. We prove (4.47): if $z \geqslant 0$, since $x \leqslant 0$ we have $z^2\,\mathrm{e}^{xz} \leqslant z^2 \leqslant a^2$, while, if $z \leqslant 0$, using $z^2 \leqslant 16\mathrm{e}^{-2-z/2}$ in this case, we obtain $z^2\,\mathrm{e}^{xz} \leqslant 16\mathrm{e}^{-2}\,\mathrm{e}^{(x-1/2)z} \leqslant 16\mathrm{e}^{-2}\mathrm{e}^{-z}$ as $x \geqslant -1/2$.

Upper bounds on the minimum in (4.46). We rewrite

$$x\,(u - \mathbb{E}[Z_1]) - x^2\,\gamma/2 = \frac{\gamma x}{2}\left(x - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right)$$

and deal with a second-order polynomial with roots 0 and $2\big(u - \mathbb{E}[Z_1]\big)/\gamma < 0$ and whose minimum over the entire real line $(-\infty, +\infty)$ is thus achieved at the midpoint $x^\star = \big(u - \mathbb{E}[Z_1]\big)/\gamma < 0$ between these roots. But the expression above is to be minimized over $[-1/2,\, 0]$ only. In the case where $u > \mathbb{E}[Z_1] - \gamma/2$, then $x^\star$ belongs to the interval of interest and

$$\min_{x \in [-1/2, 0]} \Big\{ x\,\big(u - \mathbb{E}[Z_1]\big) - x^2\,\gamma/2 \Big\} = \frac{\gamma x^\star}{2}\left(x^\star - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right) = \frac{\big(u - \mathbb{E}[Z_1]\big)^2}{2\gamma}$$

Otherwise, $u - \mathbb{E}[Z_1] \leqslant -\gamma/2$ and the midpoint $x^\star$ is to the left of $-1/2$ and the considered expression is decreasing with x on $[-1/2, 0]$, so that the minimum is achieved at $-1/2$, that is,

$$\min_{x \in [-1/2, 0]} \Big\{ x\,\big(u - \mathbb{E}[Z_1]\big) - x^2\,\gamma/2 \Big\} = -\frac{u - \mathbb{E}[Z_1]}{2} - \frac{\gamma}{8} \geqslant \frac{\gamma}{8}$$

which concludes the proof. □

4.7.2 A simplified proof of the regret bounds for MOSS and MOSS anytime

The regret bounds proven here are not new all, see Audibert and Bubeck [2009] and Degenne and Perchet [2016] for, respectively, the case of a known horizon T and the anytime version of MOSS; however the proof exposed here is somewhat simpler and more direct than in these references. In previous works, attempts were made to simultaneously build the distribution-free and some type of distribution-dependent bounds. This raised technical difficulties because of the correlations between the choices of the arms and the observed rewards. The idea of this proof is to focus solely on the distribution-free regime,

for which we notice that some crude boundings neglecting the correlations suffice (i.e., our analysis deals with all suboptimal arms in the same way, independently of how often they are played). We have also simplified the use of the peeling trick, by performing it only once on integrated quantities (instead of performing a different doubling trick for each deviation). All in all, our proof therefore consists entirely of fairly elementary and natural steps, with Hoeffding's maximal inequality in its integrated version (Corollary 4.5.2) as the only necessary technical ingredient.

To emphasize the similarity of the proofs in the anytime and non-anytime case, we present both of them in a unified fashion. The indexes used only differ by the replacement of T by t in the logarithmic exploration term in case T is unknown, see (4.4) and (4.26): compare

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\left(\frac{T}{KN_a(t)}\right)} \quad \text{and} \quad U_a^{\text{M-A}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\left(\frac{t}{KN_a(t)}\right)}$$

Note in particular that $U_a^{\text{M-A}}(t) \leqslant U_a^{\mathrm{M}}(t)$ for all arms a and all steps $1 \leqslant t \leqslant T$. We will denote by

$$U_{a,\tau}^{\mathrm{GM}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \log_+\left(\frac{\tau_t}{KN_a(t)}\right)}$$

the index of generic MOSS (GM) strategy, so that $U_a^{\mathrm{M}}(t) = U_{a,T}^{\mathrm{GM}}(t)$ and $U_a^{\text{M-A}}(t) = U_{a,t}^{\mathrm{GM}}(t)$. This GM strategy considers a sequence $(\tau_1, \dots, \tau_T)$ of integers, either $\tau_t \equiv T$ for MOSS or $\tau_t = t$ for MOSS anytime, and pick at each step $t \geqslant K+1$, an arm A_t with maximal index $U_{a,\tau}^{\mathrm{GM}}(t)$.

of Proposition 4.5.3 and of the claim after it. The first step is standard, see Bubeck and Liu [2013]. Using the fact that $U_{a^\star,\tau_t}^{\mathrm{GM}}(t) \leqslant U_{A_t^{\mathrm{GM}},\tau_t}^{\mathrm{GM}}(t)$ by definition of the index policy, the regret is smaller than

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t^{\mathrm{GM}}}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U_{a^\star,\tau_t}^{\mathrm{GM}}(t)\big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[U_{A_t^{\mathrm{GM}},\tau_t}^{\mathrm{GM}}(t) - \mu_{A_t^{\mathrm{GM}}}\Big] \tag{4.48}$$

Since $x \leqslant \delta + (x-\delta)^+$ for all x and δ, for the first inequality, by optional skipping (Section 4.5.1) for the second inequality, where we also use that pairs (a,n) such $A_t^{\mathrm{GM}} = a$ and $N_a(t) = n$ correspond to at most one $t \in \{K+1, \dots, T\}$, and by using that $U_{a,\tau}^{\mathrm{GM}}(t)$ is increasing with τ for the third inequality,

$$\begin{aligned}
\sum_{t=K+1}^{T} \mathbb{E}\Big[U_{A_t^{\mathrm{GM}}}^{\mathrm{GM}}(t) - \mu_{A_t^{\mathrm{GM}}}\Big] &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U_{A_t^{\mathrm{GM}},\tau_t}^{\mathrm{GM}}(t) - \mu_{A_t^{\mathrm{GM}}} - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U_{a,\tau_t,n}^{\mathrm{GM}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]
\end{aligned}$$

While this latter inequality may seem very crude, it turns out it is sharp enough to obtain the claimed distribution-free bounds. Moreover, it gets rid of the bothersome dependencies among the arms that are contained in the choice A_t^{GM}. Substituting in (4.48), we have shown the first inequality of Proposition 4.5.3, namely,

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big] + \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\big)^+\Big] \tag{4.49}$$

This inequality actually holds for all choices of sequences $(\tau_t)_{t\leqslant T}$ with $\tau_t \leqslant T$. The first sum in the right-hand side of (4.49) depends on the specific value of $(\tau_t)_{t\leqslant T}$ but the second sum only depends on the bound T.

Control of the left deviations of the best arm, that is, of the first sum in (4.49). For each given round $t \geqslant K+1$, we decompose

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big] = \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) < \tau_t/K\}}\Big] + \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) \geqslant \tau_t/K\}}\Big]$$

The two pieces are handled differently. The second one is easily treated by optional skipping (Section 4.5.1) and by Corollary 4.5.2, using that $U^{\mathrm{GM}}_{a^\star,\tau_t}(t) \geqslant \widehat{\mu}_{a^\star}(t)$, which actually holds with equality given $N_{a^\star}(t) \geqslant \tau_t/K$:

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) \geqslant \tau_t/K\}}\Big] \leqslant \mathbb{E}\Big[\max_{n \geqslant \tau_t/K} \big(\mu^\star - \widehat{\mu}_{a^\star,n}\big)^+\Big] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}} \tag{4.50}$$

When the arm has not been pulled often enough, we resort to a "peeling trick". We consider a real number $\beta > 1$ and further decompose the event $\{N_{a^\star}(t) < \tau_t/K\}$ along the geometric grid $x_\ell = \beta^{-\ell}\,\tau_t$, where $\ell = 0, 1, 2, \ldots$ (the endpoints x_ℓ are not necessarily integers, and some intervals $[x_{\ell+1}, x_\ell)$ may contain no integer, but none of these facts is an issue):

$$\begin{aligned}
\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t) < \tau_t/K\}}\Big] &\leqslant \sum_{\ell=0}^{+\infty} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big) \mathbb{1}_{\{x_{\ell+1} \leqslant N_{a^\star}(t) < x_\ell\}}\Big] \\
&\leqslant \sum_{\ell=0}^{+\infty} \mathbb{E}\Big[\max_{x_{\ell+1} \leqslant n < x_\ell} \big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)\Big]
\end{aligned}$$

where in the second inequality, we applied optional skipping (Section 4.5.1) once again. Now for any ℓ, the summand can be controlled as follows, first by using $n < x_\ell$ and

second by Corollary 4.5.2:

$$\begin{aligned}
\mathbb{E}\Big[\max_{x_{\ell+1}\leqslant n<x_\ell}\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)^+\Big] &= \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell}\left(\mu^\star - \widehat{\mu}_{a^\star,\tau_t,n} - \sqrt{\frac{1}{2n}\log\left(\frac{\tau_t}{Kn}\right)}\right)^+\right] \\
&\leqslant \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell}\left(\mu^\star - \widehat{\mu}_{a^\star,n} - \sqrt{\frac{1}{2x_\ell}\log\left(\frac{\tau_t}{Kx_\ell}\right)}\right)^+\right] \\
&\leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\exp\left(-\frac{x_{\ell+1}}{x_\ell}\log\left(\frac{\tau_t}{x_\ell}\right)\right) \\
&= \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\big(\beta^{-\ell}\big)^{1/\beta} = \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{\tau_t}}\,\beta^{1/2+\ell(1/2-1/\beta)}\,.
\end{aligned}$$

The above series is summable whenever $\beta \in (1,2)$. For instance we may choose $\beta = 3/2$, for which

$$\sum_{\ell=0}^{+\infty}\left(\frac{3}{2}\right)^{1/2+\ell(1/2-2/3)} = \sqrt{\frac{3}{2}}\sum_{\ell=0}^{+\infty}\alpha^{-\ell} = \frac{1}{1-\alpha}\sqrt{\frac{3}{2}} \leqslant 19 \qquad \text{where} \quad \alpha = \left(\frac{3}{2}\right)^{(1/2-2/3)}.$$

Therefore we have shown that

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big] \leqslant 19\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,. \tag{4.51}$$

Combining this bound with (4.50) and summing over t, we proved

$$\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{8}}\sum_{t=K+1}^{T}\sqrt{\frac{K}{\tau_t}}\,. \tag{4.52}$$

Control of the right deviations of all arms, that is, of the second sum in (4.49). As $(x+y)^+ \leqslant x^+ + y^+$ for all real numbers x, y, we have, for all a and $n \geqslant 1$,

$$\begin{aligned}
\big(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\big)^+ &\leqslant \big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+ + \sqrt{\frac{1}{2n}\log_+\left(\frac{T}{Kn}\right)} \\
&= \big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+ + \begin{cases} 0 & \text{if } n \geqslant T/K \\ \sqrt{\dfrac{1}{2n}\log\left(\dfrac{T}{Kn}\right)} & \text{if } n < T/K \end{cases}
\end{aligned}$$

Therefore, for each arm a,

$$\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sum_{n=1}^{T}\mathbb{E}\Big[\big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big] + \sum_{n=1}^{\lfloor T/K\rfloor}\sqrt{\frac{1}{2n}\log\left(\frac{T}{Kn}\right)} \tag{4.53}$$

We are left with two pieces to deal with separately. For the first sum in (4.53), we exploit the integrated version of Hoeffding's inequality (Corollary 4.5.2),

$$\sum_{n=1}^{T}\mathbb{E}\Big[\big(\widehat{\mu}_{a,n}-\mu_a-\sqrt{K/T}\big)^+\Big]\leqslant\sqrt{\frac{\pi}{8}}\sum_{n=1}^{T}\sqrt{\frac{1}{n}}\mathrm{e}^{-2n\left(\sqrt{K/T}\right)^2}\leqslant\sqrt{\frac{\pi}{8}}\int_0^T\sqrt{\frac{1}{x}}\,\mathrm{e}^{-2xK/T}\mathrm{d}x$$
$$=\sqrt{\frac{\pi}{8}}\sqrt{\frac{T}{2K}}\int_0^{+\infty}\frac{\mathrm{e}^{-u}}{\sqrt{u}}\mathrm{d}u=\frac{\pi}{4}\sqrt{\frac{T}{K}}\,, \tag{4.54}$$

where we used the equalities $\displaystyle\int_0^{+\infty}\big(\mathrm{e}^{-u}/\sqrt{u}\big)\mathrm{d}u=\int_0^{+\infty}\mathrm{e}^{-v^2}\mathrm{d}v=\sqrt{\pi}$.

For the second sum in (4.53), we also use a sum–integral comparison: which can be handled by comparing it to an integral and performing the change of variable $u=T/(Kx)$:

$$\sum_{n=1}^{\lfloor T/K\rfloor}\sqrt{\frac{1}{2n}\log\left(\frac{T}{Kn}\right)}\leqslant\int_0^{\lfloor T/K\rfloor}\sqrt{\frac{1}{2x}\log\left(\frac{T}{Kx}\right)}\mathrm{d}x$$
$$\leqslant\sqrt{\frac{T}{2K}}\int_1^{+\infty}u^{-3/2}\sqrt{\log(u)}\mathrm{d}u=\sqrt{\pi}\sqrt{\frac{T}{K}}$$

as $\displaystyle\int_1^{+\infty}u^{-3/2}\sqrt{\log(u)}\mathrm{d}u=2\int_0^{+\infty}v^2\,\mathrm{e}^{-v^2/2}\mathrm{d}v=\sqrt{2\pi}$ by the change of variable $u=\mathrm{e}^{v^2}$.

Conclusion. Collecting all the bounds above, we showed so far

$$R_T\leqslant(K-1)+\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\Big]+\sqrt{KT}+\sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n}-\mu_a-\sqrt{K/T}\big)^+\Big]$$
$$\leqslant(K-1)+\sqrt{KT}+\underbrace{20\sqrt{\frac{\pi}{8}}}_{\leqslant 12.6}\sum_{t=K+1}^{T}\sqrt{\frac{K}{\tau_t}}+K\underbrace{\left(\frac{\pi}{4}+\sqrt{\pi}\right)}_{\leqslant 2.6}\sqrt{\frac{T}{K}}$$

In the known horizon case $\sum 1/\sqrt{\tau_t}=T/\sqrt{T}=\sqrt{T}$ and we get $R_T\leqslant(K-1)+17\sqrt{KT}$, whereas in the anytime case,

$$\sum_{t=1}^{T}1/\sqrt{\tau_t}=\sum_{t=1}^{T}1/\sqrt{t}\leqslant\int_0^T\frac{1}{\sqrt{u}}\mathrm{d}u=2\sqrt{T}\,,$$

hence $R_T\leqslant(K-1)+29\sqrt{KT}$. □

4.7.3 Bounds for KL-UCB-Switch-Anytime

As a preliminary result to the distribution-free bound, we present an analysis of MOSS-anytime with the additional exploration φ. While we could have presented this result and Proposition 4.5.3 inside a more general result, we have chosen to separate the two to improve clarity. In the following all indices are *anytime versions with exploration function* φ.

Lemma 4.7.2 (MOSS anytime with extra-exploration)**.**

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant 29\sqrt{KT} \tag{4.55}$$

Proof. We bound both sums separately. For the first one we may recycle the bound we obtained for MOSS-anytime without the extra exploration. Indeed, as $\varphi(x) \geqslant \log_+(x)$

$$U_{a^\star}^{\text{M-A}}(t) \geqslant \widehat{\mu}_a(t) + \sqrt{\frac{1}{N_a(t)} \log_+\left(\frac{t}{KN_a(t)}\right)}$$

which is the usual MOSS-anytime index. Therefore by extracting (4.52) from the previous proof

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{2}}\sqrt{KT}$$

For the second sum we use once again the fact that the exploration vanishes at $N_a(t) \geqslant T/K$ and to bound for all arms a as in Appendix 4.7.2, eq. (4.53)

$$\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\right)^+\right] + \sum_{n=1}^{\lfloor T/K \rfloor} \sqrt{\frac{1}{n}\varphi\left(\frac{T}{Kn}\right)} \tag{4.56}$$

From (4.54) we recall that the first sum is smaller than $\pi/4\sqrt{T/K}$. The second sum is treated as before by comparison to an integral

$$\sum_{n=1}^{\lfloor T/K \rfloor} \sqrt{\frac{1}{2n}\varphi\left(\frac{T}{Kn}\right)} \leqslant \int_0^{T/K} \sqrt{\frac{1}{2x}\varphi\left(\frac{T}{Kx}\right)}\,\mathrm{d}x = \sqrt{\frac{T}{2K}} \int_1^{+\infty} \sqrt{u^{-3}\log\big(u(1+\log^2(u))\big)}\,\mathrm{d}u$$

This integral is smaller than 4. We conclude by summing over a. □

We now have all elements to provide a very short proof (with references to other results in this chapter) of the distribution-free anytime bound.

Proof of Theorem 4.2.4. Once again we begin with now usual boundings by distinguishing the value of the index depending on $N_a(t)$ for all t

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \underbrace{\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<f(t,K)\}}\Big]}_{\text{we show below that } \leqslant 8\sqrt{K/t} \text{ for each } t}$$
$$+ \underbrace{\sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t)\big)^+\Big] + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,n,T}^{\text{M-A}} - \mu_a - \sqrt{K/T}\big)^+\Big]}_{\leqslant 30\sqrt{KT} \quad \text{by (4.55)}}$$

And we are left to bound the first sum. Now since the exploration function verifies $\varphi(x) \geqslant \log_+(x)$ we may see that the index is greater than the usual KL-UCB index. Therefore the bound from the proof of Theorem 4.2.1 can be re-derived replacing T by t

$$\mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{KL-A}}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<f(t,K)\}}\Big] \leqslant 8\sqrt{\frac{K}{t}} \tag{4.57}$$

and the bound follows since $\sum_{t=1}^{T} \sqrt{1/t} \leqslant 2\sqrt{T}$ □

The distribution-dependent anytime bound is different from the known horizon case, as we do not aim for the finer second order bound.

Proof of Theorem 4.2.5.

$$\begin{aligned}\mathbb{E}[N_a(T)] &= 1 + \sum_{t=K}^{T-1} \mathbb{P}[U_{a^\star}(t) \leqslant U_a(t) \text{ and } A_{t+1} = a] \\ &\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}[U_{a^\star}(t) \leqslant \mu^\star - \delta] + \sum_{t=K}^{T-1} \mathbb{P}[U_a(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a]\end{aligned} \tag{4.58}$$

The first sum is bounded by optional skipping, Proposition 4.5.7 and Hoeffding's maximal inequality as

$$\begin{aligned}\mathbb{P}\big[U_{a^\star}(t) \leqslant \mu^\star - \delta\big] &\leqslant \sum_{n=1}^{\lfloor t/K \rfloor} \mathbb{P}\Big[\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta) \leqslant \frac{1}{n}\varphi\Big(\frac{t}{Kn}\Big)\Big] \\ &\qquad + \mathbb{P}\big[\exists n \geqslant \lfloor t/K \rfloor + 1 \ : \ \widehat{\mu}_{a,n} \leqslant \mu^\star - \delta\big] \\ &\leqslant \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2} + \mathrm{e}^{-t\delta^2/K}\end{aligned}$$

For the sake of clarity, we delay some straightforward calculations (detailed after the proof) that lead us to

$$\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2} \leqslant \frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6} \tag{4.59}$$

Hence the first sum in (4.58) is bounded by

$$\frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6}+\frac{1}{1-\mathrm{e}^{-1}}\frac{K}{\delta^2} \tag{4.60}$$

and we are now to treat the second sum. We proceed by a fine and exhaustive decomposition of the sum thanks to optional skipping. Define the event

$$\mathcal{E}_a(n,t)=\big\{N_a(t)=n \text{ and } A_{t+1}=a\big\}$$

We will use repeatedly the fact that for all n there is most one value of t such that $\mathcal{E}_a(n,t)$ holds. A direct consequence of this fact is that for any event $\mathcal{F}(n)$ that does not depend on t

$$\sum_{n=n_0}^{n_1}\sum_{t=t_0}^{t_1}\mathbb{1}_{\{\mathcal{E}_a(n,t)\text{ and }\mathcal{F}(n)\}}=\sum_{n=n_0}^{n_1}\mathbb{1}_{\{\mathcal{F}(n)\}}\underbrace{\sum_{t=t_0}^{t_1}\mathbb{1}_{\{\mathcal{E}_a(n,t)\}}}_{\leqslant 1}\leqslant\sum_{n=n_0}^{n_1}\mathbb{1}_{\{\mathcal{F}(n)\}} \tag{4.61}$$

Then by definition of the switch index

$$\begin{aligned}\sum_{t=K}^{T-1}\mathbb{P}\big[U_a(t)\geqslant\mu^\star-\delta\text{ and }A_{t+1}=a\big]&=\sum_{t=K}^{T-1}\sum_{n=1}^{t}\mathbb{P}\big[U_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]\\&=\sum_{t=K}^{T-1}\sum_{n=1}^{f(t,K)}\mathbb{P}\big[U^{\mathrm{KL}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]+\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]\end{aligned}$$

For the first sum, we may use similar bounds as in the known horizon case, as $U^{\mathrm{KL}}_{a,n,t}\leqslant U^{\mathrm{KL}}_{a,n,T}$, and then by invoking (4.61). By using the exact same calculations as in the known horizon case, see (4.22), replacing log by φ, for $\delta^2\leqslant\gamma_\star(1-\mu^\star)^2/2$ we bound the first sum by

$$\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)-2\delta/(1-\mu^\star)}+\frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}} \tag{4.62}$$

where $\gamma_\star$ is defined in (4.36). The second sum requires a more refined treatment. Define the varying threshold

$$n_1(t)=\left\lfloor\frac{8\,\varphi(t/K)}{\Delta_a^2}\right\rfloor$$

so that for $\delta\leqslant\Delta_a/2$ and $n>n_1(t)$

$$\big\{U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\big\}\subseteq\left\{\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a-\delta-\sqrt{\frac{\varphi\big(t/K\big)}{2n_1(t)}}\right\}\subseteq\big\{\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\big\} \tag{4.63}$$

We then decompose the sum as

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{n_1(t)}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]+\sum_{t=K}^{T-1}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]$$

Our choice of $n_1(t)$, via (4.63), leads to

$$\sum_{t=K}^{T}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]\leqslant\sum_{t=K}^{T}\sum_{n=n_1(t)+1}^{t}\mathbb{P}\big[\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\text{ and }\mathcal{E}_a(n,t)\big]$$

Now the event does not depend on t anymore, and thanks to (4.61) and Hoeffding's inequality, we may bound it by

$$\sum_{n=1}^{T}\mathbb{P}\big[\widehat{\mu}_{a,n}\geqslant\mu_a+\Delta_a/4\big]\leqslant\frac{1}{1-\mathrm{e}^{-\Delta_a^2/8}}$$

The only piece that remains to be bounded is now

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)+1}^{n_1(t)}\mathbb{P}\big[U^{\mathrm{M}}_{a,n,t}\geqslant\mu^\star-\delta\text{ and }\mathcal{E}_a(n,t)\big]$$

which we will bound deterministically thanks to the events $\mathcal{E}_a(n,t)$. Indeed

$$\sum_{t=K}^{T-1}\sum_{n=f(t,K)}^{n_1(t)}\mathbb{1}_{\{\mathcal{E}_a(n,t)\}}\leqslant\sum_{t=K}^{T-1}\mathbb{1}_{\{f(t,K)\leqslant n_1(t)\}}\leqslant\min\big\{t\geqslant K:f(t,K)>n_1(t)\big\}\overset{\mathrm{def}}{=}T_0 \tag{4.64}$$

since for all t, there is at most one n such that $N_a(t)=n$: hence the inside sum is at most 1, and is trivially zero whenever $f(t,K)>n_1(t)$. T_0 is a constant that depends solely on Δ_a and K.

All in all we have shown that for all T and $\delta\leqslant\min\big(\gamma_\star(1-\mu^\star)^2/2,\Delta_a/2\big)$

$$\mathbb{E}[N_a(T)]\leqslant\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)-2\delta/(1-\mu^\star)}+\frac{C_1}{\delta^6}+C_2 \tag{4.65}$$

where C_1 and C_2 are constants that do not depend on T and δ. Therefore as $T\to\infty$ we may choose $\delta=\varphi(T/K)^{-1/7}$ which gives the claimed result, remembering that $\varphi(x)=\log(x)+o(\log(x))$. □

Proof of (4.59). This is straightforward calculations : we permute the sums and compare them to the corresponding integrals

$$\begin{aligned}
\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K\rfloor}\mathrm{e}(2n+1)e^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2}&=\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\sum_{t=Kn}^{T-1}\mathrm{e}(n+2)\mathrm{e}^{-\varphi(t/(Kn))}\mathrm{e}^{-2n\delta^2}\\
&=\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\sum_{t=Kn}^{T-1}\mathrm{e}^{-\varphi(t/(Kn))}\\
&\leqslant\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\int_{t=Kn-1}^{T-1}\mathrm{e}^{-\varphi(t/(Kn))}\,\mathrm{d}t\\
&\leqslant\sum_{n=1}^{\lfloor(T-1)/K\rfloor}\mathrm{e}(2n+1)\mathrm{e}^{-2n\delta^2}\int_{u=1-1/(Kn)}^{T-1/(Kn)}Kn\mathrm{e}^{-\varphi(u)}\mathrm{d}u
\end{aligned}$$

Now we have chosen φ so that for all n

$$\int_{u=1-1/(Kn)}^{T-1/(Kn)} \mathrm{e}^{-\varphi(u)}\mathrm{d}u \leqslant \int_{1/2}^{+\infty} \mathrm{e}^{-\varphi(u)}\mathrm{d}u = \frac{1}{2} + \int_{1}^{+\infty} \frac{\mathrm{d}u}{u\big(1+\log^2(u)\big)} = \frac{1+\pi}{2} \tag{4.66}$$

Hence our sum is smaller than

$$K\mathrm{e}\frac{1+\pi}{2}\sum_{n=1}^{\lfloor (T-1)/K \rfloor} n(2n+1)\mathrm{e}^{-2n\delta^2} \leqslant \frac{5\mathrm{e}(1+\pi)}{2(1-\mathrm{e}^{-2})^3}\frac{K}{\delta^6}$$

as already detailed in (4.14). □

4.7.4 Proofs of the other results of Section 4.5.4

Proposition 4.5.7 of Section 4.5.4 was already proved in Appendix 4.7.1. We now prove the three remaining results of Section 4.5.4, namely, Lemmas 4.5.4 and 4.5.5, as well as Proposition 4.5.6.

Proof of Lemma 4.5.5

The proof of Honda and Takemura [2015, Theorem 2, Lemma 6] relies on the exhibiting the formula of interest for finitely supported distributions, via KKT conditions, and then taking limits to cover the case of all distributions. We propose a more direct approach. But before we do, we explain why it is natural to expect to rewrite $\mathcal{K}_{\mathrm{inf}}$, which is an infimum, as a maximum. Indeed, given that Kullback-Leibler divergences are given by a supremum, $\mathcal{K}_{\mathrm{inf}}$ appears as an inf sup, which under some conditions (this is Sion's lemma) is equal to a sup inf.

More precisely, a variational formula for the Kullback-Leibler divergence, see Boucheron et al. [2013, Chapter 4], has it that

$$\mathrm{KL}(\nu,\nu') = \sup\Big\{\mathbb{E}_\nu[Y] - \log \mathbb{E}_{\nu'}\big[\mathrm{e}^Y\big] \,:\, Y \text{ s.t. } \mathbb{E}_\nu[\mathrm{e}^Y] < +\infty\Big\}$$

where we indexed the expectations with respect to the underlying probability. In particular, denoting by X the identity and considering, for $\lambda \in [0,1]$, the bounded variables

$$Y_\lambda = \log\left(1 - \lambda\frac{X-\mu}{1-\mu}\right) \leqslant \log\left(1 + \frac{\lambda\mu}{1-\mu}\right)$$

we have, for any probability measure ν' such that $\mathrm{E}(\nu') > \mu$:

$$\log \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big] = \log\left(\mathbb{E}_{\nu'}\left[1 - \lambda\frac{X-\mu}{1-\mu}\right]\right) = \log\left(1 - \lambda\frac{\mathrm{E}(\nu')-\mu}{1-\mu}\right) \leqslant 0$$

Hence, for these distributions ν',

$$\mathrm{KL}(\nu,\nu') \geqslant \max_{\lambda\in[0,1]} \mathbb{E}_\nu[Y_\lambda] - \log \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big] \geqslant \max_{\lambda\in[0,1]} \mathbb{E}_\nu\left[\log\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right]$$

and by taking the infimum over all distributions ν' with $\mathrm{E}(\nu') > \mu$:

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \max_{0\leqslant\lambda\leqslant 1} \mathbb{E}_\nu\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right] \tag{4.67}$$

We now only need to prove the converse inequality.

To do so, we define the function

$$H : \lambda \in [0,1] \longmapsto \mathbb{E}_\nu\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right]$$

The function is well defined, except maybe at $\lambda = 1$ when $\nu\{1\} > 0$; we then take it equal to $-\infty$. We begin by a study of the function H.

Lemma 4.7.3. *Assume here that $\mu < \mathrm{E}(\nu) < 1$. The function H is twice differentiable on $(0,1)$ and its derivative can be defined at 1. For all $\lambda \in (0,1]$,*

$$H'(\lambda) = \frac{1}{\lambda}\left(1 - \mathbb{E}_\nu\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right) \tag{4.68}$$

Moreover, for $\lambda^\star \in \operatorname*{arg\,max}_{0\leqslant\lambda\leqslant 1} H(\lambda)$, we have

$$\mathbb{E}_\nu\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = 1 \quad \textit{if } \lambda^\star < 1 \qquad \textit{and} \qquad \mathbb{E}_\nu\left[\frac{1-\mu}{1-X}\right] \leqslant 1 \quad \textit{if } \lambda^\star = 1\,;$$

in the case when $\lambda^\star = 1$, we have in particular $\nu\{1\} = 0$.

Proof. For $\lambda \in (0,1)$, we get, by legitimately differentiating under the expectation,

$$H'(\lambda) = \mathbb{E}_\nu\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right] \quad \text{and} \quad H''(\lambda) = -\frac{1}{(1-\mu)^2}\mathbb{E}_\nu\left[\frac{(X-\mu)^2}{\left(1-\lambda\frac{X-\mu}{1-\mu}\right)^2}\right]. \tag{4.69}$$

Indeed as long as $\lambda < 1$, both variables in the expectations are bounded and we may invoke a standard differentiation theorem under the integral sign. This proves that $H'' < 0$ and therefore that H is strictly concave on $(0,1)$. Furthermore, H is continuous on $[0,1]$, possibly by defining $H(1) = -\infty$, as by monotone convergence

$$\lim_{\lambda\to 1}\mathbb{E}\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X<\mu\}}\right] = \mathbb{E}\left[\log\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X<\mu\}}\right]$$
$$\lim_{\lambda\to 1}\mathbb{E}\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X\geqslant\mu\}}\right] = \mathbb{E}\left[\log\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X\geqslant\mu\}}\right]$$

where the first expectation is finite (but the second may equal $-\infty$). The same argument shows that H' is continuous on $[0,1]$, and therefore (by a theorem on the limit of the derivatives) that H is right-derivable at 0 with derivative $-(\mathbb{E}(\nu)-\mu)/(1-\mu) > 0$. Since H is strictly concave on $(0,1)$ and continuous, it reaches its maximum exactly once in $[0,1]$. The last disjunction comes from the fact that since $H'(0) > 0$ and H' is decreasing, either $H'(1) \geqslant 0$ and H reaches its maximum at 1, or $H'(1) < 0$ and H reaches its maximum inside $(0,1)$. Since H is continuously differentiable, the derivative at the maximum is 0 in that case, which implies the equality of the expectation. □

We may now turn to the rest of the proof of Lemma 4.5.5.

Proof. For the inequality converse to (4.67), it is enough to show that there exists one value of λ and one measure ν' such that $\mathbb{E}(\nu') > \mu$ and $\nu \ll \nu'$ and

$$\mathrm{KL}(\nu,\nu') \leqslant \mathbb{E}_\nu\left[\log\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\right] \tag{4.70}$$

Recalling the definition of the KL, it thus suffices to find λ and ν' that satisfy the above conditions and

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = 1-\lambda\frac{x-\mu}{1-\mu} \qquad \nu\text{-a.s.} \tag{4.71}$$

We look for these by setting for $\lambda \in [0,1]$ the measure ν_λ defined by

$$\mathrm{d}\nu_\lambda = \frac{1}{1-\lambda\frac{x-\mu}{1-\mu}}\,\mathrm{d}\nu + \left(1-\mathbb{E}_\nu\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right)\mathrm{d}\delta_1 \tag{4.72}$$

where δ_1 is the Dirac delta measure at 1. This defines a probability measure if and only if the coefficient in front of $\mathrm{d}\delta_1$ is non-negative, i.e. if

$$\lambda H'(\lambda) \geqslant 0$$

Then for λ satifsfying this condition, ν_λ is a probability measure and $\nu \ll \nu_\lambda$. Furthermore, by (4.68):

$$\begin{aligned}
\mathrm{E}(\nu_\lambda) &= \int \frac{x}{1-\lambda(x-\mu)/(1-\mu)}\mathrm{d}\nu(x) + \lambda H'(\lambda) \\
&= \int \frac{x-\mu}{1-\lambda(x-\mu)/(1-\mu)}\mathrm{d}\nu(x) + \mu\big(1-\lambda H'(\lambda)\big) + \lambda H'(\lambda) \\
&= \mu - (1-\mu)H'(\lambda)(1-\lambda)
\end{aligned}$$

We wish to consider the case where $\mathrm{E}(\nu_\lambda) \geqslant \mu$ to use it to prove our inequality. The only value of λ that satisfies at the same time $H'(\lambda) \geqslant 0$ and $H'(\lambda)(1-\lambda) \leqslant 0$ is $\lambda^\star$, at which H reaches its maximum.

Now all that is left to prove is that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_{\lambda^\star}}(x) = 1 - \lambda^\star \frac{x-\mu}{1-\mu} \qquad \nu\text{-a.s.}$$

We do so by distinguishing two cases. If $\lambda^\star < 1$, then by Lemma 4.7.3 the expectation in (4.72) is equal to 1, that is, the $\mathrm{d}\delta_1$ comes with a 0 factor. Hence, $\nu_{\lambda^\star}$ is absolutely continuous with respect to ν, with a positive density given by the inverse of what we read in (4.72).

If $\lambda^\star = 1$, then again by Lemma 4.7.3, we know that ν does not put any probability mass at 1, which guarantees once again the desired equality. □

Proof of Lemma 4.5.4

The proof below is variations on the proofs that can be found in Honda and Takemura [2015] or earlier references.

Proof. To prove (4.29) we upper bound $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. Let a probability distribution $\nu' \in \mathcal{P}[0,1]$ be such that

$$\mathrm{E}(\nu') > \mu - \varepsilon \quad \text{and} \quad \nu' \gg \nu\,.$$

Since ν' has a countable number of atoms, one can choose a real number $x > \mu$, arbitrary close to 1, such that $\delta_x \perp \nu'$, where δ_x is the Dirac distribution at x. Let the probability distribution ν'_α be the convex combination

$$\nu'_\alpha = \alpha\delta_x + (1-\alpha)\nu'$$

where,

$$\alpha = \frac{\varepsilon}{x - (\mu - \varepsilon)}\,,$$

this choice of α entails that:

$$\mathrm{E}(\nu'_\alpha) = (1-\alpha)\,\mathrm{E}(\nu') + \alpha x > (1-\alpha)(\mu - \varepsilon) + \alpha x = \mu\,.$$

Moreover, since $\nu'_\alpha \gg \nu' \gg \nu$ and $\delta_x \perp \nu'$, one obtains the following relations between the Radon-Nikodym derivative of ν over ν' and ν'_α:

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha} = \frac{1}{1-\alpha}\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\,.$$

This allows to compute explicitly the Kullback-Leibler divergence

$$\mathrm{KL}(\nu, \nu'_\alpha) = \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\right)\mathrm{d}\nu = \mathrm{KL}(\nu, \nu') + \log\frac{1}{1-\alpha}\,.$$

Since $\mathrm{E}(\nu'_\alpha) > \mu$ and by the definition of $\mathcal{K}_{\inf}$ we can lower bound the first term in the equality above

$$\mathcal{K}_{\inf}(\nu, \mu) \leqslant \mathrm{KL}(\nu, \nu') + \log\frac{1}{1-\alpha}\,,$$

letting x go to 1, which implies α go to $\varepsilon/(1-\mu+\varepsilon)$ we have

$$\mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathrm{KL}(\nu,\nu') + \log\frac{1-\mu+\varepsilon}{1-\mu} = \mathrm{KL}(\nu,\nu') + \log\left(1+\frac{\varepsilon}{1-\mu}\right) \leqslant \mathrm{KL}(\nu,\nu') + \frac{\varepsilon}{1-\mu}$$

and thus taking the infimum over all the probability distributions ν' such that $\mathrm{E}(\nu') > \mu-\varepsilon$ entails that

$$\mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + \frac{\varepsilon}{1-\mu}\,.$$

To prove the second part (4.30), we follow the same path as above. Let a probability distribution $\nu' \in \mathcal{P}[0,1]$ be such that

$$\mathrm{E}(\nu') > \mu \quad \text{and} \quad \nu' \gg \nu\,.$$

Let the probability distribution ν'_α be the convex combination $\nu'_\alpha = (1-\alpha)\nu' + \alpha\nu$, where

$$\alpha = \frac{\varepsilon}{\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)} \in (0,1) \text{ because } \mathrm{E}(\nu) < \mu - \varepsilon\,.$$

By definition, we have $\mathrm{E}(\nu'_\alpha) = \mathrm{E}(\nu') - \alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)$, therefore $\mathrm{E}(\nu'_\alpha) > \mu - \varepsilon$. Thanks to the following order of absolute continuity $\nu' \gg \nu'_\alpha \gg \nu$, we can easily compute the Radon-Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\frac{\mathrm{d}\nu'_\alpha}{\mathrm{d}\nu} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big((1-\alpha) + \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\,,$$

and the Kullback-Leibler divergence between ν and ν':

$$\begin{aligned}
\mathrm{KL}(\nu,\nu') &= \int \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\mathrm{d}\nu + \int \log\Big((1-\alpha) + \alpha\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\mathrm{d}\nu \\
&\geqslant \int \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\mathrm{d}\nu + \alpha\int \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\mathrm{d}\nu \\
&= \mathrm{KL}(\nu,\nu'_\alpha) + \alpha\mathrm{KL}(\nu,\nu')\,.
\end{aligned}$$

where we use the concavity of logarithm. Now to recover the term $\mathcal{K}_{\inf}(\nu,\mu-\varepsilon)$ we use in this order: the Pinsker inequality, the fact that $\mathrm{KL}(\nu,\nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon)$ and $\mathrm{E}(\nu') - \mathrm{E}(\nu) \geqslant \varepsilon$,

$$\begin{aligned}
\mathrm{KL}(\nu,\nu') &\geqslant \mathrm{KL}(\nu,\nu'_\alpha) + \alpha\mathrm{KL}(\nu,\nu') \\
&\geqslant \mathrm{KL}(\nu,\nu'_\alpha) + 2\alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)^2 \\
&\geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big) \\
&\geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon^2\,.
\end{aligned}$$

To conclude it remains to take the infimum in the last inequality over the probability distributions ν' such that $\mathrm{E}(\nu') > \mu$.

□

Proof of Proposition 4.5.6

The following proof is exactly the same as that of Cappé et al. [2013, Lemma 6], except that we correct a small mistake in the constant.

Proof. Fix a real number $\gamma \in (0,1)$ and let S_γ be the set

$$S_\gamma = \left\{ \frac{1}{2} - \left\lfloor \frac{1}{2\gamma} \right\rfloor \gamma, \dots, \frac{1}{2} - \gamma, \frac{1}{2}, \frac{1}{2} + \gamma, \dots, \frac{1}{2} + \left\lfloor \frac{1}{2\gamma} \right\rfloor \gamma \right\},$$

which has at most $1 + 1/\gamma$ elements. Thanks to Lemma 4.7.4 below, for all $\widetilde{\lambda} \in [0,1]$ there exists a $\widetilde{\lambda}' \in S_\gamma$ such that for all $x \in [0,1]$

$$\log\left(1 - \widetilde{\lambda}\frac{x - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right) \leqslant 2\gamma + \log\left(1 - \widetilde{\lambda}'\frac{x - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right),$$

$$\begin{aligned} \mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) &= \max_{0 \leqslant \widetilde{\lambda} \leqslant 1} \frac{1}{n}\sum_{k=1}^{n} \log\left(1 - \widetilde{\lambda}'\frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \\ &\leqslant 2\gamma + \max_{\widetilde{\lambda} \in S_\gamma} \frac{1}{n}\sum_{k=1}^{n} \log\left(1 - \widetilde{\lambda}'\frac{X_k - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right), \end{aligned} \tag{4.73}$$

thanks to the variational representation of $\mathcal{K}_{\inf}$ (Lemma 4.5.5). It remains to apply the Markov's inequality and the union bound. Using the upper bound in Lemma 4.5.5 and the union bound we obtain

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \sum_{\widetilde{\lambda} \in S_\gamma} \mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} \log\left(1 - \widetilde{\lambda}'\frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \geqslant u - 2\gamma\right], \tag{4.74}$$

By Markov's inequality, for all $\widetilde{\lambda} \in [0,1]$ we have

$$\begin{aligned} \mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} \log\left(1 - \widetilde{\lambda}'\frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \geqslant u - 2\gamma\right] &\leqslant \mathrm{e}^{-n(u-\gamma)}\, \mathbb{E}\left[\prod_{k=1}^{n}\left(1 - \widetilde{\lambda}'\frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right)\right] \\ &= \mathrm{e}^{-n(u-2\gamma)}, \end{aligned}$$

using the independence of the X_k, thus plugging it in (4.74), we obtain

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] \leqslant \sum_{\widetilde{\lambda} \in S_\gamma} \mathrm{e}^{-n(u-2\gamma)} \leqslant (1 + 1/\gamma)\, \mathrm{e}^{-n(u-2\gamma)}$$

since the cardinality of S_γ is at most $1 + 1/\gamma$. Taking $\gamma = 1/(2n)$ allows us to conclude. □

The proof above relied on the following lemma, which is extracted from Cappé et al. [2013, Lemma 7] Its elementary proof consists in bounding of derivative of $\lambda \mapsto \log(1-\lambda c)$ and using a convexity argument.

Lemma 4.7.4. *For all $\lambda, \lambda' \in [0,1)$ such that either $\lambda \leqslant \lambda' \leqslant 1/2$ or $1/2 \leqslant \lambda' \leqslant \lambda$, for all real numbers $c \leqslant 1$,*

$$\log(1-\lambda c) - \log(1-\lambda' c) \leqslant 2|\lambda - \lambda'|$$

Proof. Note that $\psi_c : \lambda \to \log(1-\lambda c)$ is concave over $[0,1]$ and differentiable over $[0,1)$. By the concavity of ψ_c, if $\lambda < \lambda' \leqslant 1/2$, we have

$$\frac{\psi_c(\lambda) - \psi_c(\lambda')}{\lambda - \lambda'} \geqslant \psi_c'(1/2) \geqslant -2\,,$$

and if $1/2 \leqslant \lambda' < \lambda$

$$\frac{\psi_c(\lambda) - \psi_c(\lambda')}{\lambda - \lambda'} \leqslant \psi_c'(1/2) \leqslant 2\,,$$

since $c \leqslant 1$ and

$$\psi_c'(1/2) = \frac{-c}{1-c/2} = 2\frac{1-\frac{c}{2}-1}{1-\frac{c}{2}} = 2 - \frac{1}{1-\frac{c}{2}}\,.$$

The conclusion is straightforward. □

Chapter 5

Thresholding Bandit for Dose-ranging: The Impact of Monotonicity

In collaboration with Aurélien Garivier and Laurent Rossi.

Contents

5.1 Introduction

The phase 1 of clinical trials is devoted to the testing of a drug on healthy volunteers for *dose-ranging*. The first goal is to determine the maximum tolerable dose (MTD), that is the maximum amount of the drug that can be given to a person before adverse effects become intolerable or dangerous. A target tolerance level is chosen (typically 33%), and the trials aim at identifying quickly which is the dose entailing the toxicity coming closest to this level. Classical approaches are based on dose escalation, and the most well-known is the "traditional 3+3 Design": see Le Tourneau et al. [2009], Genovese et al. [2013] for and references therein for an introduction.

We propose in this chapter a complexity analysis for a simple model of phase 1 trials, which captures the essence of this problem. We assume that the possible doses are $x_1 < \ldots < x_K$, for some positive integer K. The patients are treated in sequential order, and identified by their rank. When the patient number t is assigned a dose x_k, we observe a measure of toxicity $X_{k,t}$ which is assumed to be an independent random variable. Its distribution ν_k characterizes the toxicity level of dose x_k. To avoid obfuscating technicalities, we treat here the case of Gaussian laws with known variance and unknown mean, but some results can easily be extended to other one-parameter exponential families such as Bernoulli distributions. The goal of the experiment is to identify as soon as possible the dose x_k which has the toxicity level μ_k closest to the target admissibility level S, with a controlled risk δ to make an error.

Content. This setting is an instance of the *thresholding bandit problem*: we refer to Locatelli et al. [2016] for an important contribution and a nice introduction in the fixed budget setting. Contrary to previous work, we focus here on identifying the *exact sample complexity* of the problem: we want to understand precisely (with the correct multiplicative constant) how many samples are necessary to take a decision at risk δ. We prove a lower bound which holds for all possible algorithms, and we propose an algorithm which matches this bound asymptotically when the risk δ tends to 0.

But the classical thresholding bandit problem does not catch a key feature of phase 1 clinical trials: the fact that the toxicity is *known in hindsight* to be *increasing* with the assigned dose. In other words, we investigate how many samples can be spared by algorithms using the fact that $\mu_1 < \mu_2 < \ldots < \mu_K$. Under this assumption, we prove another lower bound on the sample complexity, and provide an algorithm matching it. The sample complexity does not take a simple form (like a sum of inverse squares), but identifying it *exactly* is essential even in practice, since it is the *only way known so far* to construct an algorithm which reaches the lower bound.

We are thus able to quantify, for each problem, how many samples can be spared when means are sorted, at the cost of a slight increase in the computation cost of the algorithm.

Connections to the State of the Art. Phase 1 clinical trials have been an intense field of research in the statistical community (see Le Tourneau et al. [2009] and references

therein), but not considered as a sequential decision problem using the tools of the bandit literature. The important progress made in the recent years in the understanding of bandit models has made it possible to shed a new light on this issue, and to suggest very innovative solutions. The closest contribution are the works of Locatelli et al. [2016] and Chen et al. [2014], which provides a general framework for combinatorial pure exploration bandit problems. This work tackles the more specific issue of phase 1 trials. It aims at providing strong foundations for such solutions: it does not yet tackle all the ethical and practical constraints. Observe that it might also be relevant to look for the highest dose with toxicity *below* the target level, but practitioners do not consider this alternative goal in priority.

From a technical point the view, the approach followed here extends the theory of Best-Arm Identification initiated by Kaufmann et al. [2016] to a different setting. Building on the mathematical tools of that paper, we analyze the *characteristic time* of a thresholding bandit problem with and without the assumptions that the means are increasing. Computing the complexity with such a structural constraint on the means is a challenging task that had never been done before. It induces significant difficulties in the theory, but (by using isotonic regression) we are still able to provide a simple algorithm for computing the complexity term, which is of fundamental importance in the implementation of the algorithm. The computational complexity of the resulting algorithm is discussed in Section 5.3.1.

Organization. These lower bounds are presented in Section 5.2. We compare the complexities of the non-monotonic case versus the increasing case. This comparison is particularly simple and enlightening when $K = 2$, a setting often referred to as *A/B testing*. We discuss this case in Section 5.2.1, which furnishes a gentle introduction to the general case. We present in Section 5.3 an algorithm and show that it is asymptotically optimal when the risk δ goes to 0. The implementation of this algorithm requires, in the increasing case, an involved optimization which relies on constraint sub-gradient ascent and *unimodal regression*: this is detailed in Section 5.3.1. Section 5.3.2 shows the results of some numerical experiments for different strategies with high level of risk that complement the theoretical results. Section 5.4 summarizes further possible developments, and precedes most of the technical proofs which are given in appendix.

5.1.1 Notation and Setting

For $K \geqslant 2$, we consider a Gaussian bandit model $\big(\mathcal{N}(\mu_1, 1), \dots, \mathcal{N}(\mu_K, 1)\big)$, which we unambiguously refer to by the vector of means $\boldsymbol{\mu} = \big(\mu_1, \dots, \mu_K\big)$. Let $\mathbb{P}_{\boldsymbol{\mu}}$ and $\mathbb{E}_{\boldsymbol{\mu}}$ be respectively the probability and the expectation under the Gaussian bandit model $\boldsymbol{\mu}$. A threshold $S \in \mathbb{R}$ is given, and we denote by $a^*_{\boldsymbol{\mu}} \in \arg\min_{1 \leqslant a \leqslant K} |\mu_a - S|$ any optimal arm.

Let $\mathcal{M}$ be the set of Gaussian bandit models with an unique optimal arm and $\mathcal{I} = \{\boldsymbol{\mu} \in \mathcal{M} : \mu_1 < \dots < \mu_K\}$ be the subset of models with increasing means.

Definition of a δ-correct algorithm. A risk level $\delta \in (0,1)$ is fixed. At each step $t \in \mathbb{N}^*$ an agent chooses an arm $A_t \in \{1,\dots,K\}$ and receives a conditionally independent reward $Y_t \sim \mathcal{N}(\mu_{A_t},1)$. Let $\mathcal{F}_t = \sigma(A_1,Y_1,\dots,A_t,Y_t)$ be the information available to the player at step t. Her goal is to identify the optimal arm $a^*_{\boldsymbol{\mu}}$ while minimizing the number of draws τ. To this aim, the agent needs:

- a **sampling rule** $(A_t)_{t\geqslant 1}$, where A_t is $\mathcal{F}_{t-1}$-measurable,
- a **stopping rule** τ_δ, which is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geqslant 1}$,
- a $\mathcal{F}_{\tau_\delta}$-measurable **decision rule** $\widehat{a}_{\tau_\delta}$.

For any setting $\mathcal{S} \in \{\mathcal{M},\mathcal{I}\}$ (the non-monotonic or the increasing case), an algorithm is said to be δ-correct on $\mathcal{S}$ if for all $\boldsymbol{\mu} \in \mathcal{S}$ it holds that $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < +\infty) = 1$ and $\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_{\tau_\delta} \neq a^*_{\boldsymbol{\mu}}) \leqslant \delta$.

5.2 Lower Bounds

For $\mathcal{S} \in \{\mathcal{M},\mathcal{I}\}$, we define the set of *alternative bandit problems* of the bandit problem $\boldsymbol{\mu} \in \mathcal{M}$ by

$$\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S}) := \{\boldsymbol{\lambda} \in \mathcal{S} : a^*_{\boldsymbol{\lambda}} \neq a^*_{\boldsymbol{\mu}}\}\,, \tag{5.1}$$

and the probability simplex of dimension $K-1$ by Σ_K. The first result of this chapter is a lower bound on the sample complexity of the thresholding bandit problem, which we show in the sequel to be tight when δ is small enough.

Theorem 5.2.1. *Let $\mathcal{S} \in \{\mathcal{M},\mathcal{I}\}$ and $\delta \in (0,1/2]$. For all δ-correct algorithm on $\mathcal{S}$ and for all bandit models $\boldsymbol{\mu} \in \mathcal{S}$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geqslant T^*_{\mathcal{S}}(\boldsymbol{\mu})\,\mathrm{kl}(\delta,1-\delta)\,, \tag{5.2}$$

where the characteristic time $T^*_{\mathcal{S}}(\boldsymbol{\mu})$ is given by

$$T^*_{\mathcal{S}}(\boldsymbol{\mu})^{-1} = \sup_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^K \omega_a \frac{(\mu_a-\lambda_a)^2}{2}\,. \tag{5.3}$$

In particular, this implies that

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geqslant T^*_{\mathcal{S}}(\boldsymbol{\mu})\,.$$

This result is a generalization of Theorem 1 of Garivier and Kaufmann [2016]: the classical Best Arm Identification problem is a particular case of our non-monotonic setting $\mathcal{S} = \mathcal{M}$ with an infinite threshold $S = +\infty$. It is proved along the same lines. As Garivier and Kaufmann [2016], one proves that the supremum and the infimum are reached at a unique value, and in the sequel we denote by $\omega^*(\boldsymbol{\mu})$ the optimal weights

$$\omega^*(\boldsymbol{\mu}) := \operatorname*{arg\,max}_{\omega\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^K \omega_a \frac{(\mu_a-\lambda_a)^2}{2}\,. \tag{5.4}$$

5.2.1 The Two-armed Bandit Case

As a warm-up, we treat in the section the case $K = 2$. Here (only), one can find an explicit formula for the characteristic times.

Proposition 5.2.2. *When $K = 2$,*

$$T^*_{\mathcal{I}}(\boldsymbol{\mu})^{-1} = \frac{(2S - \mu_1 - \mu_2)^2}{8}, \tag{5.5}$$

$$T^*_{\mathcal{M}}(\boldsymbol{\mu})^{-1} = \frac{\min\big((2S - \mu_1 - \mu_2)^2, (\mu_1 - \mu_2)^2\big)}{8}. \tag{5.6}$$

Proof. The Equality (5.6) is a simple consequence of Lemma 5.2.3 proved in Section 5.5.1. It remains to treat the first Equality (5.5). Let $\boldsymbol{\mu} \in \mathcal{I}$ and suppose, without loss of generality, that arm 2 is optimal. Let $m = (\mu_1 + \mu_2)/2$ be the mean of two arms and $\Delta = \mu_2 - \mu_1$ be the gap. Noting that

$$\{\text{arm 1 is optimal}\} \Leftrightarrow m > S \quad \text{and}$$
$$\{\text{arm 2 optimal}\} \Leftrightarrow m < S\,,$$

we obtain

$$\begin{aligned} T^*_{\mathcal{I}}(\mu)^{-1} &= \sup_{\omega\in[0,1]} \inf_{\left\{\mu'_1<\mu'_2,\, |S-\mu'_1|<|S-\mu'_2|\right\}} \frac{\omega}{2}(\mu_1 - \mu'_1)^2 + \frac{1-\omega}{2}(\mu_2 - \mu'_2)^2 \\ &= \sup_{\omega\in[0,1]} A(\omega)\,, \end{aligned}$$

where $m' = (\mu'_1 + \mu'_2)/2$, $\Delta' = \mu'_2 - \mu'_1$ and we denote by $A(\omega)$ the function

$$A(\omega) := \inf_{\{\Delta'>0,\, m'>S\}} \frac{\omega}{2}\big(m - m' - (\Delta - \Delta')/2\big)^2 + \frac{1-\omega}{2}\big(m - m' + (\Delta - \Delta')/2\big)^2.$$

Writing $\chi = S - m$, easy computations lead to

$$A(\omega) = \begin{cases} 2\omega(1-\omega)\chi^2 & \text{if } \Delta + 2(2\omega - 1)\chi > 0\,, \\ \big(\chi^2 + (\Delta/2)^2 + (2\omega - 1)\chi\Delta\big)/2 & \text{else.} \end{cases}$$

Thus, since the maximum of A is attained at $\omega = 1/2$, we just proved that $T^*_{\mathcal{I}}(\mu)^{-1} = \chi^2/2$. □

Note that for both alternative sets the optimal weights defined in Equation (5.4) are uniform: $\omega^* = [1/2, 1/2]$. If the alternative set is $\mathcal{I}$, the optimal alternative, i.e. the element $\boldsymbol{\lambda}$ of $\overline{\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})}$ (the closure of $\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})$) which reaches the infimum in (5.3) for the optimal weights ω^*, is $\boldsymbol{\lambda} = [S - (\mu_2 - \mu_1)/2, S + (\mu_2 - \mu_1)/2]$. In words, in the optimal alternative the arms are translated in such a way that the mean of the two mean values is moved to the threshold S. If the alternative set is $\mathcal{M}$ and $\boldsymbol{\mu} \in \mathcal{I}$, the optimal

alternatives can be of two different forms. If the threshold is between the two mean values, then the optimal alternative is the same as for the increasing case. Otherwise, the optimal alternative is identical to the one of Best Arm Identification (see Garivier and Kaufmann [2016]): $\boldsymbol{\lambda} = [(\mu_1 + \mu_2)/2, (\mu_1 + \mu_2)/2]$. Thus, if $\mu_1 \leqslant S \leqslant \mu_2$, the two characteristic times coincide, as can be seen in Figure 5.1.

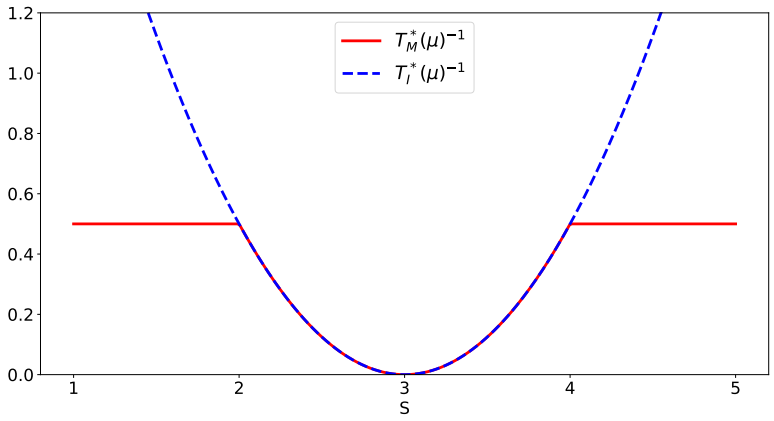

Figure 5.1: Inverse of the characteristic times as a function of the threshold S, for $\boldsymbol{\mu} = [2, 4]$. Solid red: general thresholding case ($\mathcal{S} = \mathcal{M}$). Dotted blue: increasing case ($\mathcal{S} = \mathcal{I}$).

5.2.2 On the Characteristic Time and the Optimal Proportions

We now illustrate, compare and comment the different complexities for a general bandit model $\boldsymbol{\mu} \in \mathcal{I}$ with $K \geqslant 2$ (see Figure 5.2). Since $\mathcal{I} \subset \mathcal{M}$, it is obvious that $T^*_{\mathcal{I}}(\boldsymbol{\mu}) \leqslant T^*_{\mathcal{M}}(\boldsymbol{\mu})$. The difference $T^*_{\mathcal{M}}(\boldsymbol{\mu}) - T^*_{\mathcal{I}}(\boldsymbol{\mu})$ is almost everywhere positive, and can be very large. Both $T^*_{\mathcal{I}}(\boldsymbol{\mu})$ and $T^*_{\mathcal{M}}(\boldsymbol{\mu})$ tend to $+\infty$ as S tends to middle of two consecutive arms.

On the structure of the optimal weights in the non-monotonic case.

Lemma 5.2.3. *For all $\boldsymbol{\mu} \in \mathcal{M}$,*

$$T^*_{\mathcal{M}}(\boldsymbol{\mu})^{-1} = \max_{\omega \in \Sigma_K} \min_{b \neq a^*} \frac{\omega_{a^*_{\boldsymbol{\mu}}} \omega_b}{2(\omega_{a^*_{\boldsymbol{\mu}}} + \omega_b)} \min\left((\mu_{a^*_{\boldsymbol{\mu}}} - \mu_b)^2, (2S - \mu_{a^*_{\boldsymbol{\mu}}} - \mu_b)^2\right).$$

In the non-monotonic case $\mathcal{S} = \mathcal{M}$, there are two types of optimal alternatives (as in Section 5.2.1). Indeed, the proof of Lemma 5.2.3 in Appendix 5.5.1 shows that the best alternative takes one of the two following forms. Either the optimal arm $\mu_{a^*_{\boldsymbol{\mu}}}$ and its challenger μ_b are moved to a pondered mean (by the optimal weights ω^*) of the two arms (just like in the Best Arm Identification problem), leading to a constant $(\mu_{a^*_{\boldsymbol{\mu}}} - \mu_b)^2$ in Equation (5.7). Or, as in the increasing case $\mathcal{S} = \mathcal{I}$ (see the proof of Proposition 5.2.2),

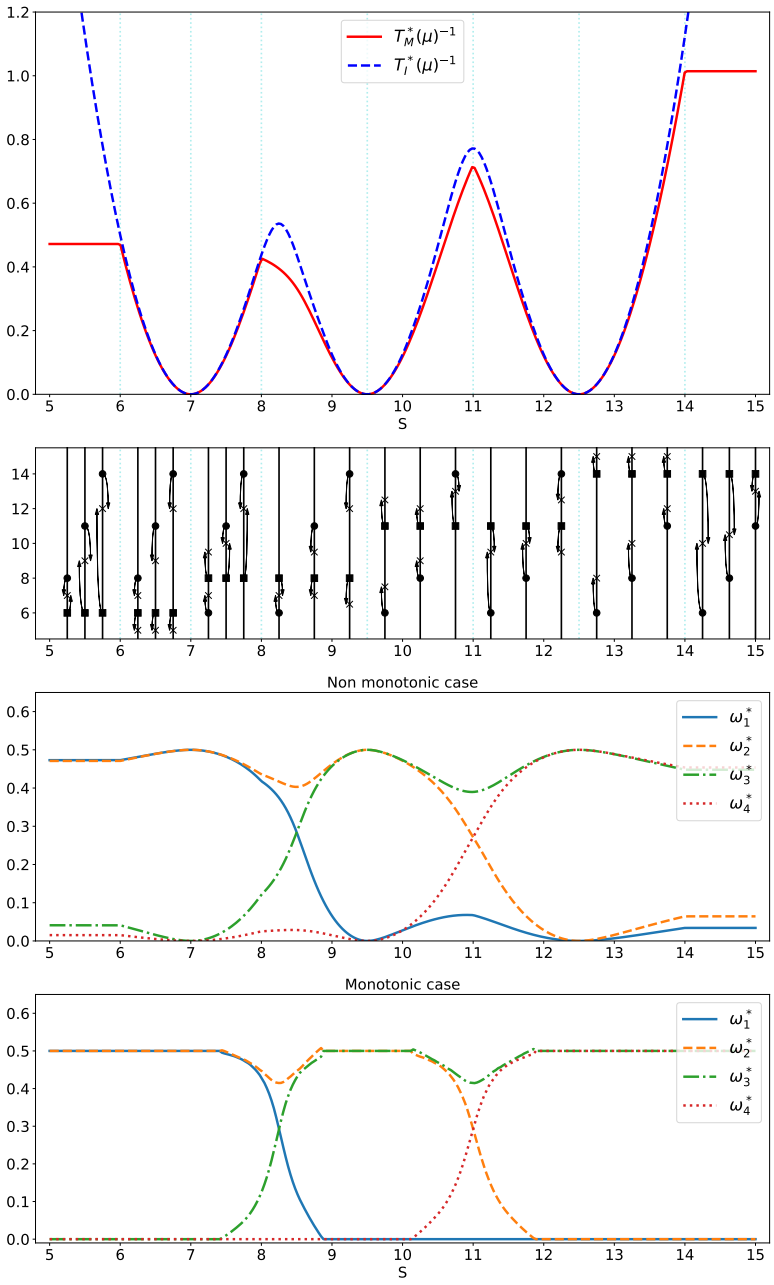

Figure 5.2: The complexity terms in the bandit model $\boldsymbol{\mu} = (6, 8, 11, 14)$. *Top:* inverse of the characteristic time as a function of the threshold S; red solid line: non-monotonic case $\mathcal{S} = \mathcal{M}$; blue dotted line: increasing case $\mathcal{S} = \mathcal{I}$. *Middle:* how to move the means to get from the initial bandit model to the optimal alternative in $\mathcal{M}$. *Bottom:* the optimal weights in function of the threshold S.

both arms $\mu_{a^*_{\boldsymbol{\mu}}}$ and μ_b are translated in the same direction, leading to the constant $(2S - \mu_{a^*_{\boldsymbol{\mu}}} - \mu_b)^2$. Figure 5.2 summarizes the different possibilities on a simple example with $K = 4$ arms, for different values of the threshold S. According to the value of S, the best alternative is shown in the second plot from the top.

On the structure of the optimal weights in the increasing case. In the increasing case $\mathcal{S} = \mathcal{I}$, one can show the remarkable property that *the optimal weights $\omega^*(\boldsymbol{\mu})$ put mass only on the optimal arm and its two closest arms.* This strongly contrasts with the non-monotonic case, as illustrated at the bottom of Figure 5.2. For simplicity we assume that $1 < a^*_{\boldsymbol{\mu}} < K$. Let $\widetilde{\omega}$ be some weights in Σ_3. Let $D^+(\theta, \widetilde{\omega})$ be the cost, with weights $\widetilde{\omega}$, for moving from the initial bandit problem $\boldsymbol{\mu}$ to a bandit problem $\widetilde{\boldsymbol{\lambda}}^+$ where arm $a^*_{\boldsymbol{\mu}}$ has mean $\theta \leqslant S$ and S is halfway between $\mu_{a^*_{\boldsymbol{\mu}}}$ and $\mu_{a^*_{\boldsymbol{\mu}}+1}$,

$$\widetilde{\lambda}^+_a = \begin{cases} \mu_a & \text{if } a > a^*_{\boldsymbol{\mu}} + 1\,, \\ 2S - \theta & \text{if } a = a^*_{\boldsymbol{\mu}} + 1\,, \\ \theta & \text{if } a = a^*_{\boldsymbol{\mu}}\,, \\ \min(\theta, \mu_a) & \text{if } a \leqslant a^*_{\boldsymbol{\mu}} - 1\,. \end{cases}$$

The explicit formula for $D^+(\theta, \widetilde{\omega})$ is

$$D^+(\theta, \widetilde{\omega}) = \widetilde{\omega}_{-1} \frac{\big(\mu_{a^*_{\boldsymbol{\mu}}-1} - \min(\mu_{a^*_{\boldsymbol{\mu}}-1}, \theta)\big)^2}{2} + \widetilde{\omega}_0 \frac{(\mu_{a^*_{\boldsymbol{\mu}}} - \theta)^2}{2} + \widetilde{\omega}_1 \frac{\big(\mu_{a^*_{\boldsymbol{\mu}}+1} - (2S - \theta)\big)^2}{2}\,.$$

Similarly we can do the same with arm $a^*_{\boldsymbol{\mu}} - 1$: moving from $\boldsymbol{\mu}$ to a bandit problem $\widetilde{\boldsymbol{\lambda}}^-$, defined for $\theta \geqslant S$ by

$$\widetilde{\lambda}^-_a = \begin{cases} \mu_a & \text{if } a < a^*_{\boldsymbol{\mu}} - 1\,, \\ 2S - \theta & \text{if } a = a^*_{\boldsymbol{\mu}} - 1\,, \\ \theta & \text{if } a = a^*_{\boldsymbol{\mu}}\,, \\ \max(\theta, \mu_a) & \text{if } a \geqslant a^*_{\boldsymbol{\mu}} + 1\,, \end{cases}$$

where both arms $a^*_{\boldsymbol{\mu}} - 1$ and $a^*_{\boldsymbol{\mu}}$ are optimal. For this alternative the cost is

$$D^-(\theta, \widetilde{\omega}) = \widetilde{\omega}_{-1} \frac{\big(\mu_{a^*_{\boldsymbol{\mu}}-1} - (2S - \theta)\big)^2}{2} + \widetilde{\omega}_0 \frac{(\mu_{a^*_{\boldsymbol{\mu}}} - \theta)^2}{2} + \widetilde{\omega}_1 \frac{\big(\mu_{a^*_{\boldsymbol{\mu}}+1} - \max(\mu_{a^*_{\boldsymbol{\mu}}+1}, \theta)\big)^2}{2}\,.$$

It appears, see the proof of Proposition 5.2.4 in Appendix 5.5.1, that these two types of alternative $\widetilde{\boldsymbol{\lambda}}^+$ and $\widetilde{\boldsymbol{\lambda}}^-$ are the optimal one. Note that they are also in $\overline{\mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})}$, the closure of the set of alternatives of $\boldsymbol{\mu}$.

Proposition 5.2.4. *For all $\boldsymbol{\mu} \in \mathcal{I}$,*

$$T^*_{\mathcal{I}}(\boldsymbol{\mu})^{-1} = \sup_{\widetilde{\omega} \in \Sigma_3} \min\Bigg(\min_{\{2S - \mu_{a^*_{\boldsymbol{\mu}}+1} \leqslant \theta \leqslant S\}} D^+(\theta, \widetilde{\omega}),\ \min_{\{S \leqslant \theta \leqslant 2S - \mu_{a^*_{\boldsymbol{\mu}}-1}\}} D^-(\theta, \widetilde{\omega}) \Bigg)\,. \tag{5.7}$$

The intuition behind this proposition is that if we try to transform $\boldsymbol{\mu}$ into an alternative $\boldsymbol{\lambda}$ with $b > a^*_{\boldsymbol{\mu}} + 1$ as optimal arm we have to pass by an alternative with optimal arm $a^*_{\boldsymbol{\mu}} + 1$ since we impose to the means to be increasing. It remains to see that this intermediate alternative has always a smaller cost. The cases with $a^*_{\boldsymbol{\mu}} = 1$ or K are similar considering only the the alternatives $\widetilde{\boldsymbol{\lambda}}^+$ if $a^*_{\boldsymbol{\mu}} = 1$ and $\widetilde{\boldsymbol{\lambda}}^-$ if $a^*_{\boldsymbol{\mu}} = K$. We can also derive bounds on the characteristic time to see that the dependence in K disappear. It is important to note that this property is really asymptotic when δ goes to zero and it is not clear at all that the dependence of the complexity in K would also disappear for moderate value of δ, we think it is not the case.

Proposition 5.2.5. *For all $\boldsymbol{\mu} \in \mathcal{I}$ such that $1 < a^*_{\boldsymbol{\mu}} < K$, considering the gaps: $\Delta^2_{-1} = (2S - \mu_{a^*_{\boldsymbol{\mu}}-1} - \mu_{a^*_{\boldsymbol{\mu}}})^2/8$, $\Delta^2_1 = (2S - \mu_{a^*_{\boldsymbol{\mu}}+1} - \mu_{a^*_{\boldsymbol{\mu}}})^2/8$ and $\Delta^2_0 = \min(\Delta^2_{-1}, \Delta^2_1)$,*

$$\frac{1}{\Delta^2_0} \leqslant T^*_{\mathcal{I}}(\boldsymbol{\mu}) \leqslant \sum_{k=-1}^{1} \frac{1}{\Delta^2_k} \leqslant \frac{3}{\Delta^2_0}\,. \tag{5.8}$$

5.3 An Asymptotically Optimal Algorithm

We present in this section an asymptotically optimal algorithm inspired by the *Direct-tracking* procedure of Garivier and Kaufmann [2016] (which borrows the idea of tracking from GAFS-MAX algorithm of Antos et al. [2008]). At any time $t \geqslant 1$ let $h(t) = (\sqrt{t} - K/2)_+$ (where $(x)_+$ stands for the positive part of x) and $U_t = \{a : N_a(t) < h(t)\}$ be the set of "abnormally rarely sampled" arms. After t rounds the empirical mean of arm a is

$$\widehat{\mu}_a(t) = \widehat{\mu}_{a,N_a(t)} = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s=a\}} ,$$

where $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{\{A_s=a\}}$ denotes the number of draws of arm a up to and including time t.

Algorithm 10: Algorithme (Direct-tracking).

Sampling rule

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}} \; N_a(t) \text{ if } U_t \neq \emptyset & (\textit{forced exploration}) \\ \underset{1 \leqslant a \leqslant K}{\operatorname{argmax}} \; t \, w_a^*(\widehat{\boldsymbol{\mu}}(t)) - N_a(t) & (\textit{direct tracking}) \end{cases}$$

Stopping rule

$$\tau_\delta = \inf \left\{ t \in \mathbb{N}^* : \widehat{\boldsymbol{\mu}}(t) \in \mathcal{M} \text{ and } \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt\left(\widehat{\boldsymbol{\mu}}(t), \mathcal{S}\right)} \sum_{a=1}^{K} N_a(t) \frac{(\widehat{\mu}_a(t) - \lambda_a)^2}{2} > \beta(t, \delta) \right\} . \tag{5.9}$$

Decision rule

$$\widehat{a}_{\tau_\delta} \in \underset{1 \leqslant a \leqslant K}{\arg\min} \left| \widehat{\mu}_a(\tau_\delta) - S \right| .$$

When $L := \operatorname{Card}\{\arg\min_{1 \leqslant a \leqslant K} |\widehat{\mu}_a(t) - S|\} > 1$, we adopt the convention that $T^*_{\mathcal{S}}\big(\widehat{\boldsymbol{\mu}}(t)\big)^{-1} = 0$ and

$$w_a^*(\widehat{\boldsymbol{\mu}}(t)) = \begin{cases} 1/L & \text{if } a \in \arg\min_{1 \leqslant a \leqslant K} |\widehat{\mu}_a(t) - S| , \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 5.3.1 (Asymptotic optimality)**.** *For $\mathcal{S} \in \{\mathcal{I}, \mathcal{M}\}$, for the constant C defined in Equation (5.22) of Section 4.4 and for $\beta(t, \delta) = \log(tC/\delta) + (3K + 2)\operatorname{loglog}(tC/\delta)$, Algorithm 10 is δ-correct on $\mathcal{S}$ and asymptotically optimal, i.e.*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leqslant T^*_{\mathcal{S}}(\boldsymbol{\mu}) . \tag{5.10}$$

The analysis of Algorithm 10 is the same in both the increasing case $\mathcal{S} = \mathcal{I}$ and the non-monotonic case $\mathcal{S} = \mathcal{M}$. It is deferred to Section 5.5.2. However, the practical implementations are quite specific to each case, and we detail them in the next section.

5.3.1 On the Implementation of Algorithm 10

The implementation of Algorithm 10 requires to compute efficiently the optimal weights $w^*(\boldsymbol{\mu})$ given by Equation (5.4). For the non-monotonic case $\mathcal{S} = \mathcal{M}$, one can follow the lines of Garivier and Kaufmann [2016], Section 2.2 and replace their Lemma 3 by Lemma 5.2.3 above.

In the increasing case $\mathcal{S} = \mathcal{I}$, however, implementing the algorithm is more involved. It is not sufficient to simply use Proposition 5.2.4, since $\widehat{\mu}(t)$ is not necessarily in $\mathcal{I}$. Let $\mathcal{I}_b := \{\boldsymbol{\lambda} \in \mathcal{I}, a^*_{\boldsymbol{\lambda}} = b\}$ be the set of alternatives with b as optimal arm. Noting that the function

$$\begin{aligned} F : w \mapsto & \inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \\ = & \min_{b \neq a^*_{\boldsymbol{\mu}}} \inf_{\boldsymbol{\lambda} \in \mathcal{I}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \end{aligned} \tag{5.11}$$

is concave (since it is the infimum of linear functions), one may access to its maximum by a sub-gradient ascent on the probability simplex Σ_K (see e.g. Boyd et al. [2003]). Let $\overline{\mathcal{I}}_b$ denote the closure of $\mathcal{I}_b$, and let

$$\boldsymbol{\lambda}^b := \arg\min_{\lambda \in \overline{\mathcal{I}}_b} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \tag{5.12}$$

be the argument of the second infimum in Equation (5.11). The sub-gradient of F at ω is

$$\partial F(\omega) = \operatorname*{Conv}_{b \in B_{Opt}} \left[\frac{(\mu_a - \lambda^b_a)^2}{2} \right]_{a \in \{1, \cdots, K\}},$$

where Conv denotes the convex hull operator and where B_{Opt} is the set of arms that reach the minimum in (5.11). Thus, performing the sub-gradient ascent simply requires to solve efficiently the minimization program (5.12). It appears that this problem boils down to *unimodal regression* (a problem closely related to isotonic regression, see for example Barlow et al. [1973] and Robertson et al. [1988]). Indeed, we can rewrite the set

$$\begin{aligned} \{\boldsymbol{\lambda} \in \mathcal{I} : a^*_{\boldsymbol{\lambda}} = b\} = \{\boldsymbol{\lambda} \in \mathcal{M} : \lambda_1 < \ldots < \lambda_{b-1} < \\ \min(\lambda_b, 2S - \lambda_b) \leqslant \max(\lambda_b, 2S - \lambda_b) < \lambda_{b+1} < \ldots < \lambda_K\}. \end{aligned}$$

Assume that $\mu_b \leqslant S$ (the other case is similar). Then $\lambda_b^b < S$, since λ_b and $2S - \lambda_b$ play a symmetric role in the constraints. Thus, in this case, one may only consider the set

$$\begin{aligned}\{\boldsymbol{\lambda} \in \mathcal{M} :& \lambda_1 < \ldots < \lambda_{b-1} < \lambda_b, \\ & 2S - \lambda_K < \ldots < 2S - \lambda_{b+1} < \lambda_b, \\ & \lambda_b \leqslant S\}\,.\end{aligned}$$

Let $\boldsymbol{\lambda}'$ be the new variables such that

$$\lambda'_a = \begin{cases} \lambda_a & \text{if } 1 \leqslant a \leqslant b\,, \\ 2S - \lambda_a & \text{else}\,. \end{cases} \tag{5.13}$$

Then $\boldsymbol{\lambda}^{b\prime}$ is the solution of the following minimization program

$$\boldsymbol{\lambda}^{b\prime} = \operatorname*{arg\,min}_{\substack{\lambda'_1 \leqslant \ldots \leqslant \lambda'_b \\ \lambda'_K \leqslant \ldots \leqslant \lambda'_b \\ \lambda'_b \leqslant S}} \sum_{a=1}^{K} \omega_a \frac{(\mu'_a - \lambda'_a)^2}{2}\,. \tag{5.14}$$

Thanks to Lemma 5.5.6 in Appendix 5.5.3, it holds that

$$\lambda_a^{b'} = \min(\widehat{\lambda}_a^b, S) \text{ for all } a \in \{1, \ldots, K\}\,,$$

where

$$\widehat{\boldsymbol{\lambda}}^b := \operatorname*{arg\,min}_{\substack{\lambda'_1 \leqslant \ldots \leqslant \lambda'_b \\ \lambda'_K \leqslant \ldots \leqslant \lambda'_b}} \sum_{a=1}^{K} \omega_a \frac{(\mu'_a - \lambda'_a)^2}{2}\,,$$

is the unimodal regression of $\boldsymbol{\mu}'$ with weights ω and with a mode located at b. It is efficiently computed via isotonic regressions (e.g. Frisén [1986], Geng and Shi [1990], Mureika et al. [1992]) with a computational complexity proportional to the number of arms K. From $\widehat{\boldsymbol{\lambda}}^b$, one can go back to $\boldsymbol{\lambda}^b$ by reversing Equation (5.13). Since we need to compute $\boldsymbol{\lambda}^b$ for each $b \neq a^*_{\boldsymbol{\mu}}$, the overall cost of an evaluation of the sub-gradient is proportional to K^2.

5.3.2 Numerical Experiments

Table 5.1 presents the results of a numerical experiment of an increasing thresholding bandit. In addition to Algorithm 10 (DT), we tried the Best Challenger (BC) algorithm with the finely tuned stopping rule given by (5.9). We also tried the Racing algorithm (R), with the elimination criterion of (5.9). For a description of all those algorithms, see Garivier and Kaufmann [2016] and references therein. Finally, in order to allow comparison with the state of the art, we added the sampling rule of algorithm APT (Anytime Parameter-free Thresholding algorithm) from Locatelli et al. [2016] in combination with the stopping rule (5.9). We chose to set the parameter ε of APT to be roughly equal to a tenth of the gap. It appears that the exploration function β prescribed in Theorem 5.3.1 is overly pessimistic. On the basis of our experiments, we recommend the use

of $\beta(t,\delta) = \log\big(\big(\log(t)+1\big)/\delta\big)$ instead. It does, experimentally, satisfy the δ-correctness property. For each algorithm, the final letter in Table 5.1 indicates whether the algorithm is aware ($\mathcal{I}$) or not ($\mathcal{M}$) that the means are increasing. We consider two frameworks:

	BC-$\mathcal{M}$	R-$\mathcal{M}$	DT-$\mathcal{M}$	APT-$\mathcal{M}$	$T^*_{\mathcal{M}}(\boldsymbol{\mu})\log\frac{1}{\delta}$
1	3913	3609	4119	5960	2033
2	3064	3164	3098	3672	1861
	BC-$\mathcal{I}$	R-$\mathcal{I}$	DT-$\mathcal{I}$	APT-$\mathcal{I}$	$T^*_{\mathcal{I}}(\boldsymbol{\mu})\log\frac{1}{\delta}$
1	483	494	611	1127	247
2	2959	2906	3072	3531	1842

Table 5.1: Monte-Carlo estimation (with 10000 repetitions) of the expected number of draws $\mathbb{E}[\tau_\delta]$ for Algorithm 10 and Best Challenger Algorithm in the increasing and non-monotonic cases. Two thresholding bandit problems are considered: bandit problem 1, $\boldsymbol{\mu}_1 = [0.5, 1.1, 1.2, 1.3, 1.4, 5]$ with $S_1 = 1$, and bandit problem 2, $\boldsymbol{\mu}_2 = [1, 2, 2.5]$ with $S_2 = 1.55$. The target risk is $\delta = 0.1$ (it is approximately reached in the first scenario, while in the second the frequency of errors is of order 1%).

in the first one, knowing that the means are increasing provides much information and gives a substantial edge: it permits to spare a large portion of the trials for the same level of risk. In the second, the complexities of the non-monotonic setting is very close to that of the increasing setting. We chose a value of the risk δ which is relatively high (10%), in order to illustrate that in this regime, the most important feature for efficiency is a finely tuned stopping rule. This shows that, even without an optimal sampling strategy, the stopping rule of (5.9) is a key feature of an efficient procedure. When the risk goes down to 0, however, optimality really requires a sampling rule which respects the proportions of Equation (5.4), as shown by Theorem 5.3.1. The poor performances of APT can be explained by the crude adaptation of this algorithm to the fixed confidence setting. This possibly comes from the fact that it was originally designed for the fixed budget setting and it appears that these two frameworks are fundamentally different, as argued by Carpentier and Locatelli [2016].

5.4 Conclusion

We provided a tight complexity analysis of the *dose-ranging* problem considered as a thresholding bandit problem with, and without, the assumption that the means of the arms are increasing. We proved that, surprisingly, the complexity terms can be computed almost as easily as in the best-arm identification case, despite the important constraints of our setting. We proposed a lower bound on the expected number of draws for any δ-correct algorithm and adapted the *Direct-Tracking* algorithm to asymptotically reach this lower bound. We also compared the complexities of the non-monotonic and the increasing cases, both in theory and on an illustrative example. We showed in Section 5.3.1 how to compute the optimal weights thanks to a sub-gradient ascent in the increasing case, a new and non-trivial task relying on unimodal isotonic regression. In order to

complement the theoretical results, we presented some numerical experiments involving different strategies in a regime of high risk. In fact, despite the asymptotic nature of the results presented here, the procedure proposed here appears to be the most efficient in practice *even when the number of trials implied is rather low* (which is often the case in clinical trials).

In the case where several arms are simultaneously closest to the threshold, the complexity of the problem is infinite. This suggests to extend the results presented here to the PAC setting, where the goal is to find *any ε-closest arm* with probability at least $1-\delta$. This extension, and extensions to the non-Gaussian case, are left for future investigation since they induce significant technical difficulties.

As a possibility of improvement, we can also mention the possible use of the unimodal regression algorithm of Stout [2000] in order to compute directly (5.11) with a complexity of order $O(K)$. We treated here mostly the case of Gaussian distributions with known variance. While the general form of the lower bound may easily be extended to other settings (including Bernoulli observations), the computation of the complexity terms is more involved and requires further investigations (in particular due to heteroscedasticity effects). The asymptotic optimality of Algorithm 10, however, can be extended directly. It remains important but very challenging tasks to make a tight analysis for moderate values of δ, to measure precisely the sub-optimality of Racing and Best Challenger strategies, and to develop a more simple and yet asymptotically optimal algorithm.

5.5 Elements of Proof

5.5.1 Proofs for the Lower Bounds

Expression of the Complexity in the Increasing Case

Fix $\boldsymbol{\mu} \in \mathcal{I}$ and let a^* be the optimal arm $a^* := a^*_{\boldsymbol{\mu}}$. We recall the definitions of $D^+(\theta, \widetilde{\omega})$ and $D^-(\theta, \widetilde{\omega})$ two functions defined over $\mathbb{R} \times \Sigma_3$ by

$$D^+(\theta, \widetilde{\omega}) = \widetilde{\omega}_{-1} \frac{\big(\mu_{a^*-1} - \min(\mu_{a^*-1}, \theta)\big)^2}{2} + \widetilde{\omega}_0 \frac{(\mu_{a^*} - \theta)^2}{2} + \widetilde{\omega}_1 \frac{\big(\mu_{a^*+1} - (2S - \theta)\big)^2}{2} \tag{5.15}$$

$$D^-(\theta, \widetilde{\omega}) = \widetilde{\omega}_{-1} \frac{\big(\mu_{a^*-1} - (2S - \theta)\big)^2}{2} + \widetilde{\omega}_0 \frac{(\mu_{a^*} - \theta)^2}{2} + \widetilde{\omega}_1 \frac{\big(\mu_{a^*+1} - \max(\mu_{a^*+1}, \theta)\big)^2}{2}\,, \tag{5.16}$$

if $1 < a^* < K$. Else, if $a^* = 1$ we define

$$\begin{aligned} D^+(\theta, \widetilde{\omega}) &= \widetilde{\omega}_0 \frac{(\mu_{a^*} - \theta)^2}{2} + \widetilde{\omega}_1 \frac{\big(\mu_{a^*+1} - (2S - \theta)\big)^2}{2} \\ D^-(\theta, \widetilde{\omega}) &= +\infty\,, \end{aligned}$$

and if $a^* = K$ we define

$$\begin{aligned} D^+(\theta, \widetilde{\omega}) &= +\infty \\ D^-(\theta, \widetilde{\omega}) &= \widetilde{\omega}_{-1} \frac{\big(\mu_{a^*-1} - (2S - \theta)\big)^2}{2} + \widetilde{\omega}_0 \frac{(\mu_{a^*} - \theta)^2}{2}\,. \end{aligned}$$

Proof of Proposition 5.2.4. We just treat here the case $1 < a^* < K$, the two other limit cases are very similar. We begin by proving that for all $\omega \in \Sigma_K$

$$\inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{S})} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2} = \min_{b \in \{a^*-1, a^*+1\}} \inf_{\{\boldsymbol{\lambda} \in \mathcal{I} \,:\, a^*_{\boldsymbol{\lambda}} = b\}} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,. \tag{5.17}$$

Indeed, let $\boldsymbol{\lambda} \in \mathcal{I}$ such that $a^*_{\boldsymbol{\lambda}} \notin \{a^* - 1, a^* + 1\}$. Suppose for example that $a^*_{\boldsymbol{\lambda}} < a^* - 1$. Let $\boldsymbol{\lambda}^\alpha$ be the family of bandit problems defined for $\alpha \in [0, 1]$ by

$$\boldsymbol{\lambda}^\alpha = \alpha \boldsymbol{\lambda} + (1 - \alpha)\boldsymbol{\mu}\,.$$

For all $\alpha \in [0, 1]$, we have $\boldsymbol{\lambda} \in \mathcal{I}$. For $\boldsymbol{\nu} \in \mathcal{I}$ and $a \in \{0, .., K\}$, let $m_a(\boldsymbol{\nu}) = (\nu_a + \nu_{a+1})/2$ be the average of two consecutive means with the convention $m_0(\boldsymbol{\nu}) = -\infty$ and $m_K(\boldsymbol{\nu}) = +\infty$. As in the case of two arms we have that $a^*_{\boldsymbol{\nu}} = a$ is equivalent to $m_a(\boldsymbol{\nu}) > S$ and $m_a(\boldsymbol{\nu}) < S$. Therefore we have the following inequalities

$$\begin{aligned} &m_{a^*_{\boldsymbol{\lambda}}-1}(\boldsymbol{\mu}) < m_{a^*_{\boldsymbol{\lambda}}}(\boldsymbol{\mu}) \leqslant m_{a^*-2}(\boldsymbol{\mu}) < m_{a^*-1}(\boldsymbol{\mu}) < S < m_{a^*}(\boldsymbol{\mu}) \text{ and} \\ &m_{a^*_{\boldsymbol{\lambda}}-1}(\boldsymbol{\lambda}) < S < m_{a^*_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}) \leqslant m_{a^*-2}(\boldsymbol{\lambda}) < m_{a^*-1}(\boldsymbol{\lambda}) < m_{a^*}(\boldsymbol{\lambda})\,. \end{aligned}$$

Thus, by continuity of the applications $\alpha \mapsto m_a(\boldsymbol{\lambda}^\alpha)$ there exits $\alpha_0 \in (0,1)$ such that

$$m_{a^*_{\boldsymbol{\lambda}}-1}(\boldsymbol{\lambda}^{\alpha_0}) < m_{a^*_{\boldsymbol{\lambda}}}(\boldsymbol{\lambda}^{\alpha_0}) \leqslant m_{a^*-2}(\boldsymbol{\lambda}^{\alpha_0}) < S < m_{a^*-1}(\boldsymbol{\lambda}^{\alpha_0}) < m_{a^*}(\boldsymbol{\lambda}^{\alpha_0})\,,$$

i.e. $a^*_{\boldsymbol{\lambda}^{\alpha_0}} = a^* - 1$. But $\alpha \mapsto \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a^\alpha)^2}{2}$ is an increasing function, and thus

$$\sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a^{\alpha_0})^2}{2} < \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,.$$

This holds for all λ, therefore

$$\inf_{\boldsymbol{\lambda} \in \mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \geqslant \min_{b \in \{a^*-1, a^*+1\}} \inf_{\{\boldsymbol{\lambda} \in \mathcal{I}\,:\, a^*_{\boldsymbol{\lambda}} = b\}} \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,.$$

The reverse inequality follows from the inclusion

$$\bigcup_{b \in \{a^*-1, a^*+1\}} \{\boldsymbol{\lambda} \in \mathcal{I} \;:\; a^*_{\boldsymbol{\lambda}} = b\} \subset \mathcal{A}lt(\boldsymbol{\mu}, \mathcal{I})\,.$$

Fix $\omega \in \Sigma_K$ and let $\boldsymbol{\lambda} \in \mathcal{I}$ be such that, say, $a^*_{\boldsymbol{\lambda}} = a^* + 1$ (the other case is similar). Then it implies $\lambda_{a^*} \leqslant S$ and we can suppose, without loss of generality, that $\lambda_{a^*} \geqslant 2S - \mu_{a^*+1}$ since it holds

$$\inf_{\{\boldsymbol{\lambda} \in \mathcal{I}\,:\, a^*_{\boldsymbol{\lambda}} = a^*+1\}} \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2} = \inf_{\{\boldsymbol{\lambda} \in \mathcal{I}\,:\, a^*_{\boldsymbol{\lambda}} = a^*+1,\ \lambda_{a^*} \geqslant 2S - \mu_{a^*+1}\}} \sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,. \tag{5.18}$$

Let $\widetilde{\boldsymbol{\lambda}}$ be such that

$$\widetilde{\lambda}_a = \begin{cases} \mu_a & \text{if } a > a^* + 1\,, \\ 2S - \lambda_{a^*} & \text{if } a = a^* + 1\,, \\ \lambda_{a^*} & \text{if } a = a^*\,, \\ \min(\lambda_{a^*}, \mu_a) & \text{if } a \leqslant a^* - 1\,. \end{cases}$$

By construction we have $\widetilde{\lambda} \in \overline{\{\lambda \in \mathcal{I} : a^*_{\boldsymbol{\lambda}} = a^* + 1\}}$. As $\lambda_{a^*+1} \leqslant 2S - \lambda_{a^*}$ and $\mu_{a^*+1} \geqslant 2S - \lambda_{a^*}$ hold, we get

$$(\widetilde{\lambda}_{a^*+1} - \mu_{a^*+1})^2 \leqslant (\lambda_{a^*+1} - \mu_{a^*+1})^2\,.$$

Similarly, for $a \leqslant a^* - 1$ we have thanks to the fact that $\lambda_a \leqslant \lambda_{a^*}$ the inequality

$$(\widetilde{\lambda}_{a^*+1} - \mu_{a^*+1})^2 \leqslant (\lambda_{a^*+1} - \mu_{a^*+1})^2\,.$$

Therefore, combining these two inequalities and using the definition of $\widetilde{\lambda}$, one obtains

$$\sum_{a=1}^K \omega_a \frac{(\mu_a - \lambda_a)^2}{2} \geqslant \sum_{a=1}^K \omega_a \frac{(\mu_a - \widetilde{\lambda}_a)^2}{2}\,,$$

and we can rewrite the infimum in Equation 5.18, indexing the alternative $\widetilde{\lambda}$ by θ the mean of arm a, as follows:

$$\begin{aligned}\inf_{\{\boldsymbol{\lambda}\in\mathcal{I}\,:\,a^*_{\boldsymbol{\lambda}}=a^*+1\}}\sum_{a=1}^{K}\omega_a\frac{(\mu_a-\lambda_a)^2}{2} &= \min_{2S-\mu_{a^*+1}\leqslant\theta\leqslant S}\sum_{a\leqslant a^*-1}\omega_a\frac{\left(\mu_a-\min\left(\theta,\mu_a\right)\right)^2}{2}\\ &\quad+\omega_{a^*}\frac{(\mu_{a^*}-\theta)^2}{2}+\omega_{a^*+1}\frac{(\mu_{a^*+1}-2S+\theta)^2}{2}\\ &= \min_{2S-\mu_{a^*+1}\leqslant\theta\leqslant S}\sum_{a<a^*-1}\omega_a\frac{\left(\mu_a-\min\left(\theta,\mu_a\right)\right)^2}{2}\\ &\quad+D^+(\theta,[\omega_{a^*-1},\omega_{a^*},\omega_{a^*+1}])\,.\end{aligned}\tag{5.19}$$

Similarly, if the optimal arm of the alternative is a^*-1, we get

$$\begin{aligned}\inf_{\{\boldsymbol{\lambda}\in\mathcal{I}\,:\,a^*_{\boldsymbol{\lambda}}=a^*-1\}}\sum_{a=1}^{K}\omega_a\frac{(\mu_a-\lambda_a)^2}{2} &= \min_{S\leqslant\theta\leqslant 2S-\mu_{a^*-1}}\sum_{a>a^*+1}\omega_a\frac{\left(\mu_a-\max\left(\theta,\mu_a\right)\right)^2}{2}\\ &\quad+D^-(\theta,[\omega_{a^*-1},\omega_{a^*},\omega_{a^*+1}])\,.\end{aligned}\tag{5.20}$$

Then, by noting that

$$\begin{aligned}\left(\mu_a-\max\left(\theta,\mu_a\right)\right)^2 &\leqslant \left(\mu_{a^*+1}-\max\left(\theta,\mu_{a^*+1}\right)\right)^2 &&\forall a\leqslant a^*+1\\ \left(\mu_a-\min\left(\theta,\mu_a\right)\right)^2 &\leqslant \left(\mu_{a^*-1}-\min\left(\theta,\mu_{a^*-1}\right)\right)^2 &&\forall a\leqslant a^*-1\end{aligned}$$

and by using the new weights $\widetilde{\omega}$ defined by

$$\widetilde{\omega}_a=\begin{cases}\sum_{b\leqslant a^*-1}w_b & \text{if } a=a^*-1\\ \omega_a & \text{if } a=a^*\\ \sum_{b\geqslant a^*+1}w_b & \text{if } a=a^*+1\\ 0 & \text{else}\,,\end{cases}$$

we obtain thanks to Equation (5.17) and to the fact that $\widetilde{\omega}$ depends only on ω:

$$\inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})}\sum_{a=1}^{K}\omega_a\frac{(\mu_a-\lambda_a)^2}{2}\leqslant\min\left(\min_{\{2S-\mu_{a^*+1}\leqslant\theta\leqslant S\}}D^+(\theta,\widetilde{\omega}),\min_{\{S\leqslant\theta\leqslant 2S-\mu_{a^*-1}\}}D^-(\theta,\widetilde{\omega})\right)\,,\tag{5.21}$$

where we identified $\widetilde{\omega}$ to an element of Σ_3. Taking the supremum on each side of (5.21), one obtains:

$$T^*_{\mathcal{I}}(\boldsymbol{\mu})^{-1}\leqslant\sup_{\widetilde{\omega}\in\Sigma_3}\min\left(\min_{\{2S-\mu_{a^*+1}\leqslant\theta\leqslant S\}}D^+(\theta,\widetilde{\omega}),\min_{\{S\leqslant\theta\leqslant 2S-\mu_{a^*-1}\}}D^-(\theta,\widetilde{\omega})\right)\,.$$

In order to prove the reverse inequality and thus (5.7), we just need to use (5.20), (5.19) and restrict the weight ω to have a support included in $\{a^*-1,a^*,a^*+1\}$. □

Proof of Proposition 5.2.5. We recall the definitions of the gaps: $\Delta_{-1}^2 = (2S - \mu_{a^*-1} - \mu_{a^*})^2/8$, $\Delta_1^2 = (2S - \mu_{a^*+1} - \mu_{a^*})^2/8$ and $\Delta_0^2 = \min(\Delta_{-1}^2, \Delta_1^2)$. For the lower bound we consider the particular weights $\overline{\omega} \in \Sigma_3$ defined by

$$\overline{\omega}_i = \frac{1/\Delta_i^2}{\sum_{k=-1}^{1} 1/\Delta_k^2} \qquad \text{for } -1 \leqslant i \leqslant 1\,.$$

Thanks to the Proposition 5.2.4, we know that

$$T_{\mathcal{I}}^*(\boldsymbol{\mu})^{-1} \geqslant \min\Bigg(\min_{\{2S-\mu_{a^*+1} \leqslant \theta \leqslant S\}} D^+(\theta, \overline{\omega}),\ \min_{\{S \leqslant \theta \leqslant 2S-\mu_{a^*-1}\}} D^-(\theta, \overline{\omega}) \Bigg)\,.$$

Then we can lower bound the two terms that appear in the minimum. Indeed, we have, denoting the mean $\overline{\theta} = \overline{\omega_0}\mu_{a^*} + \overline{\omega_1}(2S - \mu_{a^*+1})$,

$$\begin{aligned}
\min_{\{2S-\mu_{a^*+1} \leqslant \theta \leqslant S\}} D^+(\theta, \overline{\omega}) &\geqslant \min_{\{2S-\mu_{a^*+1} \leqslant \theta \leqslant S\}} \overline{\omega_0}\frac{(\mu_{a^*} - \theta)^2}{2} + \overline{\omega_1}\frac{\big((2S - \mu_{a^*+1}) - \theta)\big)^2}{2} \\
&= \overline{\omega_0}\frac{(\mu_{a^*} - \overline{\theta})^2}{2} + \overline{\omega_1}\frac{\big((2S - \mu_{a^*+1}) - \overline{\theta})\big)^2}{2} \\
&\geqslant \min(\overline{\omega_1}, \overline{\omega_0})\Delta_1^2 \geqslant \frac{1}{\sum_{k=-1}^{1} 1/\Delta_k^2}\,,
\end{aligned}$$

where we used the definition of the weights $\overline{\omega}$ for the last inequality and the fact that either $(\mu_{a^*} - \overline{\theta})^2/2 \geqslant \Delta_1^2$ or $\big((2S - \mu_{a^*+1}) - \overline{\theta})\big)^2/2 \geqslant \Delta_1^2$ since by definition $\overline{\theta}$ belongs to the interval with bounds $2S - \mu_{a^*+1}$ and μ_{a^*} for the one before. Similarly one can prove the same inequality with the second term:

$$\min_{\{S \leqslant \theta \leqslant 2S-\mu_{a^*-1}\}} D^-(\theta, \overline{\omega}) \geqslant \frac{1}{\sum_{k=-1}^{1} 1/\Delta_k^2}\,,$$

therefore we obtain the lower bound

$$T_{\mathcal{I}}^*(\boldsymbol{\mu})^{-1} \geqslant \frac{1}{\sum_{k=-1}^{1} 1/\Delta_k^2}\,.$$

For the upper bound we just need to choose a particular θ in order to bound one of the two terms that appears in the expression of $T_{\mathcal{I}}^*(\boldsymbol{\mu})^{-1}$. Thus, with the choice $\theta_1 = (2S - \mu_{a^*+1})/2 + \mu_{a^*}/2$, we get

$$\begin{aligned}
\min_{\{2S-\mu_{a^*+1} \leqslant \theta \leqslant S\}} D^+(\theta, \widetilde{\omega}) &\leqslant D^+(\theta_1, \widetilde{\omega}) \\
&\leqslant (\widetilde{\omega}_{-1} + \widetilde{\omega}_0 + \widetilde{\omega}_1)\frac{(2S - \mu_{a^*+1} - \mu_{a^*})^2}{8} = \Delta_1^2\,,
\end{aligned}$$

where we used that $\big(\mu_{a^*-1}-\min(\mu_{a^*-1},\theta_1)\big)^2/2 \leqslant (2S-\mu_{a^*+1}-\mu_{a^*})^2/8$ since θ_1 is at the middle between $2S-\mu_{a^*+1}$ and μ_{a^*}. In the same way with $\theta_{-1} = (2S-\mu_{a^*-1})/2+\mu_{a^*}/2$ one can obtain

$$\min_{\{2S-\mu_{a^*+1}\leqslant\theta\leqslant S\}} \leqslant \Delta_{-1}^2 \,.$$

Combining these two inequalties with Proposition 5.2.4 leads to

$$T_{\mathcal{I}}^*(\boldsymbol{\mu})^{-1} \leqslant \Delta_0^2 \,.$$

□

Expression of the Complexity in the Non-monotonic Case

Proof of Lemma 5.2.3. To simplify the notations we note $a_{\boldsymbol{\mu}}^* = a^*$. Thanks to the definition of the characteristic time, we just have to prove that

$$\inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{M})} \sum_{a=1}^K \omega_a \frac{(\mu_a-\lambda_a)^2}{2} = \min_{b\neq a^*} \frac{\omega_{a^*}\omega_b}{2(\omega_{a^*}+\omega_b)} \min\left((\mu_{a^*}-\mu_b)^2, (2S-\mu_{a^*}-\mu_b)^2\right).$$

Using that

$$\mathrm{Alt}(\boldsymbol{\mu},\mathcal{M}) = \bigcup_{b\neq a^*} \{\boldsymbol{\lambda}\in\mathcal{M} : |\lambda_b - S| < |\lambda_{a^*}-S|\},$$

one has

$$\begin{aligned}\inf_{\boldsymbol{\lambda}\in\mathcal{A}lt(\boldsymbol{\mu},\mathcal{S})} \sum_{a=1}^K \omega_a \frac{(\mu_a-\lambda_a)^2}{2} &= \min_{b\neq a^*} \inf_{|\lambda_b-S|<|\lambda_{a^*}-S|} \sum_{a=1}^K \omega_a \frac{(\mu_a-\lambda_a)^2}{2} \\ &= \min_{b\neq a^*} \inf_{|\lambda_b-S|<|\lambda_{a^*}-S|} \omega_{a^*}\frac{(\mu_{a^*}-\lambda_{a^*})^2}{2} + \omega_b\frac{(\mu_b-\lambda_b)^2}{2}.\end{aligned}$$

Since at the infimum it holds $|\lambda_b - S| = |\lambda_{a^*} - S|$, denoting $x = \lambda_b - S$, we have $\lambda_{a^*} - S = x$ or $-x$. Therefore, one obtains

$$\begin{aligned}\inf_{|\lambda_b-S|<|\lambda_{a^*}-S|} \omega_{a^*}\frac{(\mu_{a^*}-\lambda_{a^*})^2}{2} + \omega_b\frac{(\mu_b-\lambda_b)^2}{2} = \min\Bigg(&\inf_x \omega_a^*\frac{(\mu_{a^*}-S-x)^2}{2} + \omega_b\frac{(\mu_b-S-x)^2}{2}, \\ &\inf_x \omega_{a^*}\frac{(\mu_{a^*}-S+x)^2}{2} + \omega_b\frac{(\mu_b-S-x)^2}{2}\Bigg).\end{aligned}$$

Noting that

$$\begin{aligned}\inf_x \omega_{a^*}\frac{(\mu_{a^*}-S-x)^2}{2} + \omega_b\frac{(\mu_b-S-x)^2}{2} &= \frac{\omega_{a^*}\omega_b}{2(\omega_{a^*}+\omega_b)}(\mu_{a^*}-\mu_b)^2, \\ \inf_x \omega_{a^*}\frac{(\mu_{a^*}-S-x)^2}{2} + \omega_b\frac{(\mu_b-S+x)^2}{2} &= \frac{\omega_{a^*}\omega_b}{2(\omega_{a^*}+\omega_b)}(2S-\mu_{a^*}-\mu_b)^2,\end{aligned}$$

permits to conclude. □

5.5.2 Correctness and Asymptotic Optimality of Algorithm 10

Proof of Proposition 5.3.1. We follow and slightly adapt the proof of Theorem 14 of Kaufmann et al. [2016]. We fix a bandit problem $\boldsymbol{\mu} \in \mathcal{S}$ and the constant

$$C := e^{K+1} \left(\frac{2}{K}\right)^K \big(2(3K+2)\big)^{3K} \frac{4}{\log(3)} \,. \tag{5.22}$$

We begin by proving that Algorithm 10 is δ-correctness on $\mathcal{S}$, then we show that it is asymptotically optimal.

δ-correctness on $\mathcal{S}$

We will prove in the second part of proof that τ is almost surely finite, confer (5.25). By definition of τ, the probability that the predicted arm is the wrong one is upper-bounded by

$$\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_\tau \neq a^*_{\boldsymbol{\mu}}) \leqslant \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N}^*, \sum_{a=1}^K N_a(t) \frac{(\widehat{\mu}_a(t) - \mu_a)^2}{2} > \beta(t,\delta) \right) , \tag{5.23}$$

where we used that $\boldsymbol{\mu} \in \mathcal{A}lt\big(\widehat{\boldsymbol{\mu}}(t), \mathcal{S}\big)$ since $\widehat{a}_\tau \neq a^*_{\boldsymbol{\mu}}$. Using the union bound then Theorem 5.5.5 (note that $\beta(t,\delta) \geqslant K+1$ thanks to the choice of C) we have

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}(\widehat{a}_\tau \neq a^*_{\boldsymbol{\mu}}) &\leqslant \sum_{t=1}^{+\infty} \mathbb{P}_{\boldsymbol{\mu}}\left(\sum_{a=1}^K N_a(t) \frac{(\widehat{\mu}_a(t) - \mu_a)^2}{2} > \beta(t,\delta) \right) \\
&\leqslant \sum_{t=1}^{+\infty} e^{K+1} \left(\frac{2}{K}\right)^K \Big(\beta(t,\delta)\big(\log(t)\beta(t,\delta) + 1\big)\Big)^K e^{-\beta(t,\delta)} \\
&\leqslant e^{K+1} \left(\frac{2}{K}\right)^K \sum_{t=1}^{+\infty} \frac{\big(2(3K+2)\big)^{3K}}{\log(tC/\delta)^2} \frac{\delta}{tC} \\
&\leqslant e^{K+1} \left(\frac{2}{K}\right)^K \big(2(3K+2)\big)^{3K} \sum_{t=1}^{+\infty} \frac{1}{t\log(3t)^2} \frac{\delta}{C} \\
&\leqslant e^{K+1} \left(\frac{2}{K}\right)^K \big(2(3K+2)\big)^{3K} \frac{2}{\log(3)} \frac{\delta}{C} \leqslant \delta \,,
\end{aligned}$$

where in the third inequality we replaced $\beta(t,\delta)$ by its value and used in the fourth inequality, for $C \geqslant 3$, the following upper-bound

$$\sum_{t=1}^{+\infty} \frac{1}{t\log(3t)^2} \leqslant \frac{1}{\log(3)^2} + \int_{t=1}^{+\infty} \frac{1}{t\log(3t)^2} \mathrm{d}t \leqslant \frac{2}{\log(3)} \,.$$

Asymptotic Optimality

We begin by remarking that the function $\mu \to \omega^*(\mu)$ is continuous on the sets $\mathcal{S}_b = \{\boldsymbol{\mu} \in \mathcal{S} : \ a^*_{\boldsymbol{\mu}} = b\}$ for $b \in \{1, \dots, K\}$. Indeed it is a consequence of Lemma 5.2.3 if $\mathcal{S} = \mathcal{M}$ and Proposition 5.2.4 if $\mathcal{S} = \mathcal{I}$ and the Maximum theorem from Berge [1963]. Let ε be

a real in $(0,1)$. From the continuity of w^* in $\boldsymbol{\mu}$, there exists $\alpha = \alpha(\varepsilon)$ such that the neighbourhood of $\boldsymbol{\mu}$:

$$I_\varepsilon := [\mu_1 - \alpha, \mu_1 + \alpha] \times \cdots \times [\mu_K - \alpha, \mu_K + \alpha]$$

is such that for all $\boldsymbol{\mu}' \in I_\varepsilon$,

$$\boldsymbol{\mu}' \in \mathcal{S}, \quad a^*_{\boldsymbol{\mu}} = a^*_{\boldsymbol{\mu}'} \text{ and } \quad \max_a |w^*_a(\boldsymbol{\mu}') - w^*_a(\boldsymbol{\mu})| \leqslant \varepsilon\,.$$

Let $T \in \mathbb{N}^*$ and define the typical event where $\widehat{\boldsymbol{\mu}}(t)$ is not too far from $\boldsymbol{\mu}$

$$\mathcal{E}_T(\varepsilon) = \bigcap_{t=T^{1/4}}^{T} (\widehat{\boldsymbol{\mu}}(t) \in I_\varepsilon)\,.$$

The two following Lemmas are extracted from Kaufmann et al. [2016].

Lemma 5.5.1. *There exists two constants B, C (that depend on $\boldsymbol{\mu}$ and ε) such that*

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c) \leqslant BT\exp(-CT^{1/8})\,.$$

Lemma 5.5.2. *There exists a constant T_ε such that for $T \geqslant T_\varepsilon$, it holds that on $\mathcal{E}_T$,*

$$\forall t \geqslant \sqrt{T},\ \max_a \left| \frac{N_a(t)}{t} - w^*_a(\mu) \right| \leqslant 2(K-1)\varepsilon$$

We now assume that $T \geqslant T_\varepsilon$. Introducing the constant

$$C^*_\varepsilon(\boldsymbol{\mu}) = \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}|| \leqslant \alpha(\varepsilon) \\ \boldsymbol{w}' : ||\boldsymbol{w}' - \boldsymbol{w}^*(\boldsymbol{\mu})|| \leqslant 2(K-1)\varepsilon}} \inf_{\lambda \in \mathcal{A}lt\left(\boldsymbol{\mu}', \mathcal{S}\right)} \sum_{a=1}^{K} w_a \frac{(\mu'_a(t) - \lambda_a)^2}{2}\,,$$

thanks to Lemma 5.5.2, on the event $\mathcal{E}_T$ it holds that for every $t \geqslant \sqrt{T}$,

$$t \inf_{\lambda \in \mathcal{A}lt\left(\widehat{\boldsymbol{\mu}}(t), \mathcal{S}\right)} \sum_{a=1}^{K} \frac{N_a(t)}{t} \frac{(\widehat{\mu}_a(t) - \lambda_a)^2}{2} \geqslant tC^*_\varepsilon(\boldsymbol{\mu})\,. \tag{5.24}$$

Thus, combining (5.24) and the definition of the stopping rule (5.9), we have on the event $\mathcal{E}_T$

$$\begin{aligned} \max(\tau_\delta, T) &\leqslant \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(\tau_\delta > t)} \\ &\leqslant \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{\left\{tC^*_\varepsilon(\boldsymbol{\mu}) \leqslant \beta(T,\delta)\right\}} \leqslant \sqrt{T} + \frac{\beta(T,\delta)}{C^*_\varepsilon(\boldsymbol{\mu})}. \end{aligned}$$

Introducing

$$T_0(\delta) = \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{\beta(T,\delta)}{C_\varepsilon^*(\boldsymbol{\mu})} \leqslant T\right\},$$

for every $T \geqslant \max(T_0(\delta), T_\varepsilon)$, one has $\mathcal{E}_T \subseteq \{\tau_\delta \leqslant T\}$, therefore thanks to Lemma 5.5.1

$$\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta > T) \leqslant \mathbb{P}(\mathcal{E}_T^c) \leqslant BT\exp(-CT^{1/8})$$

and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \leqslant T_0(\delta) + T_\varepsilon + \sum_{T=1}^{\infty} BT\exp(-CT^{1/8}). \tag{5.25}$$

We now provide an upper bound on $T_0(\delta)$. Introducing the constant

$$H(\varepsilon) = \inf\{T \in \mathbb{N} : T - \sqrt{T} \geqslant T/(1+\varepsilon)\}$$

one has

$$\begin{aligned} T_0(\delta) &\leqslant H(\varepsilon) + \inf\left\{T \in \mathbb{N} : \beta(T,\delta) \leqslant \frac{C_\varepsilon^*(\boldsymbol{\mu})T}{1+\varepsilon}\right\} \\ &\leqslant H(\varepsilon) + \inf\left\{T \in \mathbb{N} : \log(TC/\delta) + (3K+2)\,\mathrm{loglog}(TC/\delta) \leqslant \frac{C_\varepsilon^*(\boldsymbol{\mu})T}{1+\varepsilon}\right\}. \end{aligned}$$

Using technical Lemma 5.5.4, for δ small enough to have $\big(C_\varepsilon^*(\boldsymbol{\mu})\delta\big)/\big((1+\varepsilon)^2 C\big) \leqslant e$, we get

$$\begin{aligned} T_0(\delta) &\leqslant C(\varepsilon) + \frac{\delta}{C}\max\left(g\left(\frac{C_\varepsilon^*(\boldsymbol{\mu})\delta}{(1+\varepsilon)^2 C}\right), \exp\left(g\left(\frac{\varepsilon}{3K+2}\right)\right)\right) \\ &\leqslant C(\varepsilon) + \max\left(\frac{(1+\varepsilon)^2}{C_\varepsilon^*(\boldsymbol{\mu})}\log\left(\frac{e(1+\varepsilon)^2 C}{C_\varepsilon^*(\boldsymbol{\mu})\delta}\log\left(\frac{(1+\varepsilon)^2 C}{C_\varepsilon^*(\boldsymbol{\mu})\delta}\right)\right), \frac{\delta}{C}\exp\left(g\left(\frac{\varepsilon}{3K+2}\right)\right)\right). \end{aligned}$$

This last upper bound yields, for every $\varepsilon > 0$,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leqslant \frac{(1+\varepsilon)^2}{C_\varepsilon^*(\boldsymbol{\mu})}.$$

Letting ε tend to zero and by definition of w^*,

$$\lim_{\varepsilon\to 0} C_\varepsilon^*(\boldsymbol{\mu}) = T_{\mathcal{S}}^*(\boldsymbol{\mu})^{-1},$$

allows us to conclude

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leqslant T_{\mathcal{S}}^*(\boldsymbol{\mu}).$$

□

5.5.3 Some Technical Lemmas

We regroup in this Appendix some technical lemmas used in the asymptotic analysis of Algorithm 10.

5.5.4 An Inequality

For $0 < y \leqslant 1/e$ let g be the function

$$g(y) = \frac{1}{y}\log\left(\frac{e}{y}\log\left(\frac{1}{y}\right)\right) . \tag{5.26}$$

Lemma 5.5.3. *Let $A > 0$ such that $1/A > e$, then for all $x \geqslant g(A)$*

$$\log(x) \leqslant Ax\,. \tag{5.27}$$

Proof. Since $g(A) \geqslant 1/A$, the function $x \mapsto A - 1/x$ is non-decreasing, we just need to prove (5.27) for $x = g(A)$. It remains to remark that

$$\begin{aligned}\log\big(g(A)\big) &\leqslant \log\left(\frac{2}{A}\log\left(\frac{1}{A}\right)\right)\\ &\leqslant \log\left(\frac{e}{A}\log\left(\frac{1}{A}\right)\right) = A\,g(A)\,,\end{aligned}$$

as $\log(x) \leqslant x/e$. □

Lemma 5.5.4. *Let $A, B > 0$, then for all $\varepsilon \in (0,1)$ such that $(1+\varepsilon)/A < e$ and $B/\varepsilon > e$, for all $x \geqslant \max\Big(g\big(A/(1+\varepsilon)\big), \exp\big(g(\varepsilon/B)\big)\Big)$*

$$\log(x) + B\log\log(x) \leqslant Ax\,. \tag{5.28}$$

Proof. Since $\log(x) \geqslant g(\varepsilon/B)$ thanks to Lemma 5.5.3 we have $B\log\log(x) \leqslant \varepsilon\log(x)$. Therefore , still using Lemma 5.5.3 with $x \geqslant g\big(A/(1+\varepsilon)\big)$,

$$\begin{aligned}\log(x) + B\log\log(x) &\leqslant (1+\varepsilon)\log(x)\\ &\leqslant Ax\,.\end{aligned}$$

□

A Deviation Bound

We recall here for self-containment the Theorem 2 of Magureanu et al. [2014].

Theorem 5.5.5. *For all $\delta \geqslant (K+1)$ and $t \in \mathbb{N}^*$ we have*

$$\mathbb{P}\left(\sum_{a=1}^{K} N_a(t)\frac{(\widehat{\mu}_a(t) - \mu_a)^2}{2} \geqslant \delta\right) \leqslant e^{K+1}\left(\frac{2\delta(\delta\log(t)+1)}{K}\right)^K e^{-\delta}\,. \tag{5.29}$$

The factor 2 that differs from Theorem 2 of Magureanu et al. [2014] comes from the fact that we consider deviation at the right and left of the mean.

Unimodal Regression under Bound Restriction

For $\boldsymbol{\mu} \in \mathcal{M}$, $\omega \in \mathring{\Sigma}_K$ (where $\mathring{\Sigma}_K$ stands for the interior of Σ_K) and $b \in \{1, \cdots, K\}$, let $\mathcal{U}$ be the set of unimodal vector with maximum localized at b

$$\mathcal{U} = \{\boldsymbol{\lambda} : \lambda_1 \leqslant \cdots \leqslant \lambda_b \geqslant \lambda_{b+1} \geqslant \cdots \lambda_K\}\,, \tag{5.30}$$

and $\mathcal{U}_S$ be the same set with an additional bound restriction on λ_b

$$\mathcal{U}_S = \{\boldsymbol{\lambda} : \lambda_1 \leqslant \cdots \leqslant \lambda_b \geqslant \lambda_{b+1} \geqslant \cdots \lambda_K,\ \lambda_b \leqslant S\}\,. \tag{5.31}$$

Let $\widehat{\boldsymbol{\lambda}}$ be the unimodal regression of $\boldsymbol{\mu}$

$$\widehat{\boldsymbol{\lambda}} := \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \mathcal{U}} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,, \tag{5.32}$$

and $\boldsymbol{\lambda}^*$ be the projection of $\boldsymbol{\mu}$ on $\mathcal{U}_S$

$$\boldsymbol{\lambda}^* := \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \mathcal{U}_S} \sum_{a=1}^{K} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}\,. \tag{5.33}$$

We have, as in the case of isotonic regression (see Hu [1997]), the following simple relation between $\boldsymbol{\lambda}^*$ and $\widehat{\boldsymbol{\lambda}}$

Lemma 5.5.6. *It holds that*

$$\lambda_a^* = \min(\widehat{\lambda}_a, S) \text{ for all } a \in \{1, \ldots, K\}\,.$$

To prove Lemma 5.5.6 we need the following properties on $\widehat{\boldsymbol{\lambda}}$.

Lemma 5.5.7. *Let $c_{-k} < \ldots < c_0 > \ldots > c_l$ be real numbers and $(A_{-k}, \ldots, A_0, \ldots, A_k)$ be integer intervals forming a partition of $\{1, \ldots, K\}$ be such that $\widehat{\boldsymbol{\lambda}}$ is constant on the sets A_i equals to c_i for all $-k \leqslant i \leqslant l$ and $b \in A_0$. Then, for all $-k \leqslant i \leqslant l$ and $\boldsymbol{\lambda} \in \mathcal{U}$*

$$\sum_{a \in A_i} (\mu_a - \widehat{\lambda}_a)\omega_a = 0 \tag{5.34}$$

$$\sum_{a \in A_i} (\mu_a - \widehat{\lambda}_a)\omega_a \lambda_a \leqslant 0\,. \tag{5.35}$$

Proof. Since $\widehat{\boldsymbol{\lambda}}$ is the projection of $\boldsymbol{\mu}$ on the closed convex $\mathcal{U}$ we know that for all $\boldsymbol{\lambda}$ in $\mathcal{U}$

$$\sum_{A \in \{1, \ldots, K\}} (\mu_a - \widehat{\lambda}_a)(\widehat{\lambda}_a - \lambda_a)\omega_a \geqslant 0\,. \tag{5.36}$$

Fix $\boldsymbol{\lambda} \in \mathcal{U}$ and $-k \leqslant i \leqslant l$ and suppose, for example, that $i < 0$. The other cases $i = 0$ and $i > 0$ are similar. Introduce, for $|\varepsilon| < \min\left(|c_i - c_{i-1}|, |c_{i+1} - c_i|\right)$, the vector $\boldsymbol{\lambda}^\varepsilon$ such that

$$\lambda_a^\varepsilon = \begin{cases} c_i - \varepsilon & \text{if } a \in A_i\,, \\ \widehat{\lambda}_a & \text{else.} \end{cases}$$

By construction $\boldsymbol{\lambda}^\varepsilon \in \mathcal{U}$ and thanks to (5.36) we have

$$\sum_{A\in\{1,\dots,K\}} (\mu_a - \widehat{\lambda}_a)(\widehat{\lambda}_a - \lambda_a^\varepsilon)\omega_a = \varepsilon \sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)\omega_a \geqslant 0\,.$$

Taking ε positive or negative proves (5.34). Let $x, y \in \{1,\dots,K\}$ be such that $A_i = \{x, x+1, \dots, y-1, y\}$ and $\boldsymbol{\lambda}'$ be such that

$$\lambda'_a = \begin{cases} \lambda_a & \text{if } a \in A_i\,, \\ \lambda_x & \text{if } a < x\,, \\ \lambda_y & \text{if } a > y\,. \end{cases}$$

By construction $\boldsymbol{\lambda}' \in \mathcal{U}$ and thanks to (5.36) we have

$$\begin{aligned}\sum_{A\in\{1,\dots,K\}} (\mu_a - \widehat{\lambda}_a)(\widehat{\lambda}_a - \lambda_a)\omega_a &= \sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)(c_i - \lambda'_a)\omega_a \\ &\quad + \lambda_x \sum_{j<i}\sum_{a\in A_j} (\mu_a - \widehat{\lambda}_a)\omega_a + \lambda_y \sum_{j>i}\sum_{a\in A_j} (\mu_a - \widehat{\lambda}_a)\omega_a \\ &= -\sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)\lambda'_a\omega_a\,,\end{aligned}$$

where we used (5.34). Equation (5.36) allows us to prove (5.35). □

We now adapt the proof of Hu [1997] to the case of unimodal regression.

Proof of Lemma 5.5.6. Since $\mathcal{U}_S$ is a closed convex we just need to check that for all $\lambda \in \mathcal{U}_S$

$$\sum_{a\in\{1,\dots,K\}} \big(\mu_a - \min(\widehat{\lambda}_a, S)\big)\big(\min(\widehat{\lambda}_a, S) - \lambda_a\big)\omega_a \geqslant 0\,.$$

We have, using the same notation of Lemma 5.5.7,

$$\begin{aligned}\sum_{a\in\{1,\dots,K\}} \big(\mu_a - \min(\widehat{\lambda}_a, S)\big)\big(\min(\widehat{\lambda}_a, S) - \lambda_a\big)\omega_a &= \sum_{i:c_i\leqslant S}\sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)(\widehat{\lambda}_a - \lambda_a)\omega_a \\ &\quad + \sum_{i:c_i>S}\sum_{a\in A_i} (\mu_a - S)(S - \lambda_a)\omega_a \\ &= \sum_{i:c_i\leqslant S}\sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)(\widehat{\lambda}_a - \lambda_a)\omega_a \\ &\quad + \sum_{i:c_i>S}\sum_{a\in A_i} (\mu_a - \widehat{\lambda}_a)(S - \lambda_a)\omega_a + \sum_{i:c_i>S}\sum_{a\in A_i} (c_i - S)(\lambda_a - S)\omega_a \geqslant 0\,,\end{aligned}$$

where we used the Lemma 5.5.7 for the two first sums and the fact that $\lambda_a < S$ for the last sum. □

Chapter 6

Fano's inequality for random variables

In collaboration with Sébastien Gerchinovitz and Gilles Stoltz.

Contents

6.1 Introduction

Fano's inequality is a popular information-theoretical result that provides a lower bound on worst-case error probabilities in multiple-hypotheses testing problems. It has important consequences in information theory [Cover and Thomas, 2006] and related fields. In mathematical statistics, it has become a key tool to derive lower bounds on minimax (worst-case) rates of convergence for various statistical problems such as nonparametric density estimation, regression, and classification (see, e.g., Tsybakov, 2009, Massart, 2007).

Multiple variants of Fano's inequality have been derived in the literature. They can handle a finite, countable, or even continuously infinite number of hypotheses. Depending on the community, it has been stated in various ways. In this chapter, we focus on statistical versions of Fano's inequality. For instance, its most classical version states that for all sequences of $N \geqslant 2$ probability distributions $\mathbb{P}_1, \ldots, \mathbb{P}_N$ on the same measurable space $(\Omega, \mathcal{F})$, and all events $A_1, \ldots, A_N$ forming a partition of Ω,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N}\inf\limits_{\mathbb{Q}} \displaystyle\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \log(2)}{\log(N)},$$

where the infimum in the right-hand side is over all probability distributions $\mathbb{Q}$ over $(\Omega, \mathcal{F})$. The following alternative version is popular among statisticians and is due to Birgé [2005] and Massart [2007]: with the same notation and conditions,

$$\min_{1 \leqslant i \leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{c, \frac{\overline{K}}{\log(N)}\right\} \qquad \text{where} \qquad \overline{K} = \frac{1}{N-1}\sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1) \tag{6.1}$$

for some universal constant $c \in (0, 1)$. In both cases, the link to multiple-hypotheses testing is the following: when applied to events of the form $A_i = \{\widehat{\theta} = i\}$, the last inequality provides a lower bound on the worst-case error probability $\max\limits_{1 \leqslant i \leqslant N} \mathbb{P}_i\big(\widehat{\theta} \neq i\big)$ for any estimator $\widehat{\theta}$.

Several extensions to more complex settings were derived in the past. For example, Han and Verdú [1994] addressed the case of countably infinitely many probability distributions, while Duchi and Wainwright [2013] and Chen et al. [2016] further generalized Fano's inequality to continuously infinitely many distributions. Gushchin [2003] extended Fano's inequality in another direction, by considering $[0, 1]$–valued random variables Z_i such that $Z_1 + \ldots + Z_N = 1$, instead of the special case $Z_i = \mathbb{1}_{A_i}$. All these extensions, as well as others recalled in Section 6.6, provide a variety of tools that adapt nicely to the variety of statistical problems.

Main contributions. In this chapter, we revisit Fano's inequality and make the following three sets of contributions. First, we extend Fano's inequality to both continuously many distributions $\mathbb{P}_\theta$ and arbitrary $[0, 1]$–valued random variables Z_θ that

are not required to sum up (or integrate) to 1. We also point out that the alternative distribution $\mathbb{Q}$ could vary with θ. Despite the high degree of generality, the proofs of these results are simple thanks to a reduction to Bernoulli distributions.

Second, we provide new statistical applications, illustrating in particular that it is handy to be able to consider random variables (not necessarily summing up to 1). The two main such applications deal with Bayesian posterior concentration lower bounds and a regret lower bound in non-stochastic sequential learning.

Finally, as a by-product of our simplified analysis, we highlight a direct connection between Fano's and Pinsker's inequalities. We prove a common bound that both implies Pinsker's inequality for $N = 2$ and a Fano-type inequality for all $N \geqslant 2$. These two inequalities were classically thought to be useful in distinct regimes (Pinsker's inequality for $N = 2$, Fano's inequality for $N \geqslant 3$). This is one reason why Birgé [2005] designed his alternative version (6.1) of the most classical version of Fano's inequality in order to make it nontrivial even for $N = 2$.

Content and outline of this chapter. The main body of this chapter contains new results and a new look at some older results (that we sometimes generalize), while the appendix contains omitted technical derivations and discussions, or even some known material (which we provide for the sake of self-completeness).

More precisely, Sections 6.2 and 6.3 explain our two-step methodology to obtain several versions of Fano's inequality, at various degrees of generality. These inequalities are discussed and compared to the literature later in the chapter, in Section 6.6. Before that, we present in Section 6.4 our two main applications: lower bounds for minimax Bayesian posterior concentration and for non-stochastic sequential learning. Section 6.5 presents two other applications which—perhaps surprisingly—follow from the special case $N = 1$ in Fano's inequality. One of these applications is about distribution-dependent lower bounds on Bayesian posterior concentration (elaborating on results by Hoffmann et al., 2015). Section 6.7 concludes the main body of the chapter and provides new and simpler proofs of some important bounds on the Kullback-Leibler divergence, the main contributions being a short and enlightening proof of the refined Pinsker's inequality by Ordentlich and Weinberger [2005], and a sharper Bretagnolle and Huber [1978, 1979] inequality.

The appendix of the present chapter contains the following material. In Section 6.8.1, we present two toy applications of our continuous Fano's inequality in parametric and nonparametric regression. Section 6.8.2 provides some background on the problem of Bayesian posterior concentration. Section 6.8.3 carefully discusses the popular version of Fano's inequality proved by Birgé [2005] and Massart [2007]. Section 6.8.4 is a reminder of basic properties of f–divergences (such as the data-processing inequality), and Section 6.8.5 explains how our two-step methodology readily extends to f–divergences. Finally, Section 6.8.6 states and proves a version of Jensen's inequality tailored to the needs of the present chapter: that holds for general convex sets and for possibly infinite-valued convex functions.

Notation. Let $\mathbb{P}, \mathbb{Q}$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$. We write $\mathbb{P} \ll \mathbb{Q}$ to indicate that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Moreover, the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ is defined by

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \begin{cases} \displaystyle\int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} & \text{if } \mathbb{P} \ll \mathbb{Q}; \\ +\infty & \text{otherwise.} \end{cases}$$

We write $\mathrm{Ber}(p)$ for the Bernoulli distribution with parameter p. We also use the usual measure-theoretic conventions in $\mathbb{R} \cup \{+\infty\}$; in particular $0 \times (+\infty) = 0$ and $1/0 = +\infty$, as well as $0/0 = 0$. We also set $\log(0) = -\infty$ and $0 \log(0) = 0$.

6.2 How to derive a Fano-type inequality: an example

In this section we explain on an example the methodology to derive Fano-type inequalities. We will present the generalization of the approach and the resulting bounds in Section 6.3, but the proof below already contains the two key arguments: a reduction to Bernoulli distributions, and a lower bound on the kl function. We discuss how novel (or not novel) our results and approaches are in Section 6.6.

Proposition 6.2.1. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i, \mathbb{Q}_i$ and all events A_i (non necessarily disjoint), where $i \in \{1, \dots, N\}$, with $0 < \frac{1}{N} \sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, we have*

$$\frac{1}{N} \sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{\dfrac{1}{N} \displaystyle\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i) + \log(2)}{-\log\left(\dfrac{1}{N} \displaystyle\sum_{i=1}^N \mathbb{Q}_i(A_i)\right)}.$$

In particular, if $N \geqslant 2$ and the A_i form a partition,

The proof uses the Kullback-Leibler divergence function kl between Bernoulli distributions: for all $(p, q) \in [0, 1]^2$,

$$\mathrm{kl}(p, q) \stackrel{\text{def}}{=} \mathrm{KL}\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right),$$

with the usual measure-theoretic conventions. The proof also relies on the two following information-theoretic tools, which are proved in Appendix 6.8.4. Lemma 6.2.2 indicates that transforming the data at hand can only reduce the ability to distinguish between two probability distributions. Corollary 6.2.3 states that the Kullback-Leilber divergence is jointly convex.

Lemma 6.2.2 (Contraction of entropy; also known as data-processing inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}^X$ and $\mathbb{Q}^X$ the laws of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{KL}\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q})\,.$$

Corollary 6.2.3 (Joint convexity of KL)**.** *The Kullback-Leibler divergence* KL *is jointly convex, i.e., for all probability distributions $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}_1, \mathbb{Q}_2$ over the same measurable space $(\Omega, \mathcal{F})$, and all $\lambda \in (0,1)$,*

$$\mathrm{KL}\big(\lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2,\, \lambda\mathbb{Q}_1 + (1-\lambda)\mathbb{Q}_2\big) \leqslant \lambda\,\mathrm{KL}(\mathbb{P}_1, \mathbb{Q}_1) + (1-\lambda)\,\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2)\,.$$

Proof (of Proposition 6.2.1): Our first step is to reduce the problem to Bernoulli distributions. Using first the joint convexity of the Kullback-Leibler divergence (Corollary 6.2.3), and second the data-processing inequality with the indicator functions $X = \mathbb{1}_{A_i}$ (Lemma 6.2.2), we get

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i),\, \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{6.2}$$

Therefore, we have $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$ with

$$\overline{p} = \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \qquad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) \qquad \overline{K} = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{6.3}$$

Our second and last step is to lower bound $\mathrm{kl}\big(\overline{p}, \overline{q}\big)$ to extract an upper bound on $\overline{p}$. Noting that $\overline{p}\log\big(\overline{p}\big) + \big(1-\overline{p}\big)\log\big(1-\overline{p}\big) \geqslant -\log(2)$, we have, by definition of $\mathrm{kl}\big(\overline{p}, \overline{q}\big)$,

$$\mathrm{kl}\big(\overline{p}, \overline{q}\big) \geqslant \overline{p}\log\big(1/\overline{q}\big) - \log(2)\,, \qquad \text{thus} \qquad \overline{p} \leqslant \frac{\mathrm{kl}\big(\overline{p}, \overline{q}\big) + \log(2)}{\log\big(1/\overline{q}\big)}\,. \tag{6.4}$$

Substituting the upper bound $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$ in (6.4) concludes the proof. □

6.3 Various Fano-type inequalities, with the same two ingredients

We extend the approach of Section 6.2 and derive a broad family of Fano-type inequalities, which will be of the form

$$\overline{p} \leqslant \psi\big(\overline{q}, \overline{K}\big)\,,$$

where the average quantities $\overline{p}$, $\overline{q}$ and $\overline{K}$ are described in Section 6.3.1 and where the functions ψ are described in Section 6.3.2. The simplest example that we considered in Section 6.2 was given, for some probability distributions $\mathbb{P}_i, \mathbb{Q}_i$ and some events A_i, by

$$\overline{p} = \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \qquad \overline{q} = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) \qquad \overline{K} = \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,.$$

But we also address here the more general case where the finite averages are replaced with integrals over any measurable space Θ, and where the indicator functions $\mathbb{1}_{A_i}$ are replaced with arbitrary $[0,1]$–valued random variables Z_θ, where $\theta \in \Theta$.

Section 6.3.3 states some examples of such Fano-type inequalities, based on a choice of averages picked in Section 6.3.1 and a choice of functions ψ picked in Section 6.3.2.

We recall that the novelty (or lack of novelty) of our results will be discussed in detail in Section 6.6.

6.3.1 Reduction to Bernoulli distributions

As in Section 6.2, we can use the contraction of relative entropy to lower bound any Kullback-Leibler divergence by that of suitably chosen Bernoulli distributions. We present four such reductions, in increasing degree of generality. We only recall how to prove the first one, since they are all similar.

Finitely many distributions; uniform averages. We consider some underlying measurable space, N pairs of probability distributions $\mathbb{P}_i$, $\mathbb{Q}_i$ on this space, and N events A_i, where $i \in \{1, \dots, N\}$. The events A_i do not need to be disjoint. Recall from Section 6.2 that

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i),\ \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,, \tag{6.5}$$

where the first inequality is by joint convexity of the Kullback-Leibler divergence (Corollary 6.2.3 above), and where the second inequality is by the data-processing inequality (Lemma 6.2.2 above), considering the indicator functions $X = \mathbb{1}_{A_i}$.

Countably many distributions; general averages. The argument above carries over to any convex combination $\alpha = (\alpha_1, \alpha_2, \dots)$ of countably many pairs of probability distributions $\mathbb{P}_i$, $\mathbb{Q}_i$ and events A_i, where $i \in \{1, 2, \dots\}$. The convex combination α can be thought of as a prior distribution. Let $\delta_{(x,y)}$ denote the Dirac mass at $(x, y) \in \mathbb{R}^2$. Using the general form of Jensen's inequality stated in Lemma 6.8.12 (Appendix 6.8.6) with $\varphi(p, q) = \mathrm{kl}(p, q)$ on the convex set $C = [0,1]^2$, together with the probability measure $\mu = \sum_i \alpha_i\, \delta_{\left(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\right)}$, we get

$$\mathrm{kl}\left(\sum_{i\geqslant 1} \alpha_i\, \mathbb{P}_i(A_i),\ \sum_{i\geqslant 1} \alpha_i\, \mathbb{Q}_i(A_i)\right) \leqslant \sum_{i\geqslant 1} \alpha_i \mathrm{kl}\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \sum_{i\geqslant 1} \alpha_i \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{6.6}$$

Distributions indexed by a possibly continuous set; general averages. We consider statistical models $\mathbb{P}_\theta$, $\mathbb{Q}_\theta$ with a measurable parameter space $(\Theta, \mathcal{G})$, a prior probability distribution ν over Θ, and a collection A_θ of events (not necessarily disjoint) such that

$$\theta \in \Theta \longmapsto \big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big) \qquad \text{and} \qquad \theta \in \Theta \longmapsto \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

are $\mathcal{G}$–measurable. The reduction is this time (we use again the general form of Jensen's inequality in Appendix 6.8.6, Lemma 6.8.12):

$$\begin{aligned}\mathrm{kl}\left(\int_\Theta \mathbb{P}_\theta(A_\theta)\mathrm{d}\nu(\theta),\, \int_\Theta \mathbb{Q}_\theta(A_\theta)\mathrm{d}\nu(\theta)\right) &\leqslant \int_\Theta \mathrm{kl}\big(\mathbb{P}_\theta(A_\theta),\, \mathbb{Q}_\theta(A_\theta)\big)\mathrm{d}\nu(\theta) \\ &\leqslant \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)\mathrm{d}\nu(\theta)\,. \end{aligned} \tag{6.7}$$

Random variables; general averages. In the reductions above, it was unnecessary that the sets A_i or A_θ form a partition or even be disjoint. It is therefore not surprising that the former reductions can be generalized by replacing the indicator functions $\mathbb{1}_{A_i}$ or $\mathbb{1}_{A_\theta}$ with arbitrary $[0,1]$–valued random variables Z_i or Z_θ. The most elegant way of generalizing the reduction is the following consequence of Lemma 6.2.2 (extracted from Chapter 2 and proved again in Appendix 6.8.4 for the sake of self-completeness).

Corollary 6.3.1 (Contraction of entropy; with expectations of random variables)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$ taking values in $[0,1]$. Denote by $\mathbb{E}_\mathbb{P}[X]$ and $\mathbb{E}_\mathbb{Q}[X]$ the expectations of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{kl}\big(\mathbb{E}_\mathbb{P}[X], \mathbb{E}_\mathbb{Q}[X]\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q})\,.$$

We now state the reduction in the case of finitely many distributions and uniform averages, as well as in the case of distributions indexed by a possibly continuous set. In the first case, we consider a collection $Z_1, \dots, Z_N$ of random variables taking values in $[0,1]$ and denote by $\mathbb{E}_{\mathbb{P}_i}$ and $\mathbb{E}_{\mathbb{Q}_i}$ the expectations with respect to $\mathbb{P}_i$ and $\mathbb{Q}_i$; the reduction is

$$\mathrm{kl}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{P}_i}\big[Z_i\big],\, \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big]\right) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{kl}\Big(\mathbb{E}_{\mathbb{P}_i}\big[Z_i\big],\, \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big]\Big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,. \tag{6.8}$$

In the most general case, the $[0,1]$–valued random variables are denoted by Z_θ, where $\theta \in \Theta$, and expectations with respect to $\mathbb{P}_\theta$ and $\mathbb{Q}_\theta$ are denoted by $\mathbb{E}_{\mathbb{P}_\theta}$ and $\mathbb{E}_{\mathbb{Q}_\theta}$; we assume that

$$\theta \in \Theta \longmapsto \Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\, \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\Big) \qquad \text{and} \qquad \theta \in \Theta \longmapsto \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)$$

are $\mathcal{G}$–measurable. The reduction then is

$$\begin{aligned}\mathrm{kl}\left(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta),\, \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta)\right) &\leqslant \int_\Theta \mathrm{kl}\Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\, \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\Big)\mathrm{d}\nu(\theta) \\ &\leqslant \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}_\theta)\mathrm{d}\nu(\theta)\,. \end{aligned} \tag{6.9}$$

This most general form of the reduction will be used in Section 6.4.1.

6.3.2 Any lower bound on kl leads to a Fano-type inequality

The section above indicates that after the reduction to the Bernoulli case, we get inequations of the form ($\overline{p}$ is usually the unknown)

$$\mathrm{kl}\big(\overline{p},\overline{q}\big) \leqslant \overline{K}\,,$$

where $\overline{K}$ is an average of Kullback-Leibler divergences, and $\overline{p}$ and $\overline{q}$ are averages of probabilities of events or expectations of $[0,1]$–valued random variables.

We thus proceed by lower bounding the kl function. The first bound was already used in Section 6.2.

The most classical bound. For all $p \in [0,1]$ and $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\log(1/q) - \log(2)\,, \qquad \text{thus} \qquad p \leqslant \frac{\mathrm{kl}(p,q)+\log(2)}{\log(1/q)}\,. \tag{6.10}$$

This bound can be improved by replacing the term $\log(2)$ with $\log(2-q)$, which leads to a non-trivial bound even if $q = 1/2$ (as is the case in some applications).

A consequence of a convexity inequality. This bound was known and we recall its proof in Section 6.7.1. For all $p \in [0,1]$ and $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant p\log(1/q) - \log(2-q)\,, \qquad \text{thus} \qquad p \leqslant \frac{\mathrm{kl}(p,q)+\log(2-q)}{\log(1/q)}\,. \tag{6.11}$$

A (novel) consequence of this bound is that

$$p \leqslant 0.21 + 0.79\,q + \frac{\mathrm{kl}(p,q)}{\log(1/q)}\,. \tag{6.12}$$

A final bound, of a similar flavor, is stated below. Note that, perhaps surprisingly, it makes a connection between Pinsker's and Fano's inequalities.

A consequence of a refined Pinsker's inequality. The first inequality was known, the second is a novel but straightforward consequence of it. We provide the proofs in Section 6.7.2. For all $p \in [0,1]$ and $q \in (0,1)$,

$$\mathrm{kl}(p,q) \geqslant \max\left\{\log\left(\frac{1}{q}\right), 2\right\}(p-q)^2\,, \qquad \text{thus} \qquad p \leqslant q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\big\{\log(1/q),2\big\}}}\,. \tag{6.13}$$

6.3.3 Examples of combinations

The combination of (6.8) and (6.13) ensures the following Fano-type inequality for finitely many random variables, whose sum does not need to be 1.

Lemma 6.3.2. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i$, $\mathbb{Q}_i$ and for all $[0,1]$–valued random variables Z_i defined on this measurable space, where $i \in \{1, \ldots, N\}$, with*

$$0 < \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big] < 1\,,$$

we have

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}_i}\big[Z_i\big] \leqslant \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big] + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}_i)}{-\log\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}_i}\big[Z_i\big]\right)}}\,.$$

The combination of (6.7) and (6.10) yields a continuous version of Fano's inequality. (We discard again all measurability issues.)

Lemma 6.3.3. *We consider a mesurable space $(\Theta, \mathcal{E})$ equipped with a probability distribution ν. Given an underlying measurable space $(\Omega, \mathcal{F})$, for all two collections $\mathbb{P}_\theta$, $\mathbb{Q}_\theta$, of probability distributions over this space and all collections of events A_θ of $(\Omega, \mathcal{F})$, where $\theta \in \Theta$, with*

$$0 < \int_\Theta \mathbb{Q}_\theta(A_\theta) \mathrm{d}\nu(\theta) < 1\,,$$

we have

$$\int_\Theta \mathbb{P}_\theta(A_\theta) \mathrm{d}\nu(\theta) \leqslant \frac{\displaystyle\int_\Theta \mathrm{KL}\big(\mathbb{P}_\theta, \mathbb{Q}_\theta\big) \mathrm{d}\nu(\theta) + \log(2)}{\displaystyle -\log\left(\int_\Theta \mathbb{Q}_\theta(A_\theta) \mathrm{d}\nu(\theta)\right)}\,.$$

6.3.4 Extensions to f–divergences

Gushchin [2003] generalized Fano-type inequalities with the Kullback-Leibler divergence (as above) to arbitrary f–divergences, in the case where finitely many $[0,1]$–valued random variables $Z_1 + \ldots + Z_N = 1$ are considered. As we discuss in Appendix 6.8.5, the main reason why this generalization was possible is that f–divergences also satisfy a data-processing inequality. We show that all the reductions to Bernoulli distributions discussed in Section 6.3.1 go through for f–divergences as well.

6.3.5 On the sharpness of the obtained bounds

The reductions of Section 6.3.1 are sharp in the sense that they can hold with equality (they cannot be improved at this level of generality). Now, we want to draw from the result of this first reduction, which is of the form $\mathrm{kl}\big(\overline{p}, \overline{q}\big) \leqslant \overline{K}$, an upper bound on $\overline{p}$.

We introduce the the generalized inverse of kl in its second argument: for all $q \in [0,1]$ and all $y \geqslant 0$,

$$\mathrm{kl}(\,\cdot\,, q)^{(-1)}(y) \stackrel{\text{def}}{=} \sup\bigl\{p \in [0,1] :\ \mathrm{kl}(p,q) \leqslant y\bigr\}\,;$$

when $q \in (0,1)$, it is thus equal to the largest root q of the equation $\mathrm{kl}(p,q) = y$ if $y \leqslant \log(1/q)$ or to 1 otherwise. We then get

$$\overline{p} \leqslant \mathrm{kl}\bigl(\,\cdot\,, \overline{q}\bigr)^{(-1)}\bigl(\overline{K}\bigr)\,.$$

This formulation should be reminiscent of Birgé [2005, Theorem 2], but has one major practical drawback: it is unreadable, and this is why we considered the lower bounds of Section 6.3.2.

Question is now how sharp these lower bounds on kl are. They are all (in spirit) of the form

$$\mathrm{kl}(p,q) \geqslant p\log\left(\frac{1}{q}\right) - \ldots \qquad \text{thus} \qquad p \leqslant \frac{\mathrm{kl}(p,q)}{\log(1/q)} + \ldots\,,$$

where the ... on the right refer to terms that vanish when $q \to 0$. In the applications, q is typically small and the main term $\mathrm{kl}(p,q)/\log(1/q)$ is of the order of a constant. Therefore, the lemma below explains that up to the ... terms, the bounds of Section 6.3.2 are essentially optimal.

Lemma 6.3.4. *For all $q \in (0,1)$ and $p \in [0,1]$, whenever $p \geqslant q$, we have*

$$\mathrm{kl}(p,q) \leqslant p\log\left(\frac{1}{q}\right) \qquad \textit{thus} \qquad p \geqslant \frac{\mathrm{kl}(p,q)}{\log(1/q)}\,.$$

Proof. We note that when $p \geqslant q$, we have $(1-p)/(1-q) \leqslant 1$, so that

$$\mathrm{kl}(p,q) = p\log\left(\frac{1}{q}\right) + \underbrace{p\log(p)}_{\leqslant 0} + (1-p)\underbrace{\log\left(\frac{1-p}{1-q}\right)}_{\leqslant 0} \leqslant p\log\left(\frac{1}{q}\right)\,,$$

hence the first inequality. □

6.4 Main applications

We present two new applications of Fano's inequality, with $[0,1]$–valued random variables Z_i or Z_θ. The topics covered are:

- Bayesian posterior concentration rates, for which we use the reduction (6.9);
- robust sequential learning (prediction of individual sequences) in the case of sparse losses, which relies on the reduction (6.8).

As can be seen below, the fact that we are now able to consider arbitrary $[0,1]$–valued random variables Z_θ on a continuous parameter space Θ makes the proof of the Bayesian posterior concentration lower bound quite simple.

For pedagogical purposes, we also illustrate in Appendix 6.8.1 how to use the continuous Fano's inequality for parametric or nonparametric regression. Two more applications will also be presented in Section 6.5; they have a different technical flavor, as they rely on only one pair of distributions, i.e., $N = 1$.

6.4.1 Lower bounds on Bayesian posterior concentration rates

In the next paragraphs we show how our continuous Fano's inequality can be used in a simple fashion to derive lower bounds for posterior concentration rates.

Setting and Bayesian terminology. We consider the following density estimation setting: we observe a sample of independent and identically distributed random variables $X_{1:n} = (X_1, \dots, X_n)$ drawn from a probability distribution P_θ on $(\mathcal{X}, \mathcal{F})$, with a fixed but unknown $\theta \in \Theta$. We assume that the measurable parameter space $(\Theta, \mathcal{G})$ is equipped with a prior distribution π and that all $P_{\theta'}$ have a density $p_{\theta'}$ with respect to some reference measure $\mathfrak{m}$ on $(\mathcal{X}, \mathcal{F})$. We also assume that $(x, \theta') \mapsto p_{\theta'}(x)$ is $\mathcal{F} \otimes \mathcal{G}$–measurable. We can thus consider the transition kernel $(x_{1:n}, A) \mapsto \mathbb{P}_\pi(A \,|\, x_{1:n})$ defined for all $x_{1:n} \in \mathcal{X}^n$ and all sets $A \in \mathcal{G}$ by

$$\mathbb{P}_\pi(A \,|\, x_{1:n}) = \frac{\displaystyle\int_A \prod_{i=1}^n p_{\theta'}(x_i) \mathrm{d}\pi(\theta')}{\displaystyle\int_\Theta \prod_{i=1}^n p_{\theta'}(x_i) \mathrm{d}\pi(\theta')} \tag{6.14}$$

if the denominator lies in $(0, +\infty)$; if it is null or infinite, we set, e.g., $\mathbb{P}_\pi(A \,|\, x_{1:n}) = \pi(A)$. The resulting random measure $\mathbb{P}_\pi(\,\cdot \,|\, X_{1:n})$ is known as the *posterior* distribution.

Let $\ell : \Theta \times \Theta \to \mathbb{R}_+$ be a measurable loss function that we assume to be a pseudo-metric[1]. A posterior concentration rate with respect to ℓ is a sequence $(\varepsilon_n)_{n \geqslant 1}$ of positive real numbers such that, for all $\theta \in \Theta$,

$$\mathbb{E}_\theta \Big[\mathbb{P}_\pi \big(\theta' : \ell(\theta', \theta) \leqslant \varepsilon_n \,\Big|\, X_{1:n} \big) \Big] \longrightarrow 1 \qquad \text{as } n \to +\infty \,,$$

where $\mathbb{E}_\theta$ denotes the expectation with respect to $X_{1:n}$ where each X_j has the P_θ law. The above convergence guarantee means that, as the size n of the sample increases, the posterior mass concentrates in expectation on an ε_n–neighborhood of the true parameter θ. Several variants of this definition exist (e.g., convergence in probability or almost surely; or ε_n that may depend on θ). Though most of these definitions can be handled with the techniques provided below, we only consider this one for the sake of conciseness.

[1] The only difference with a metric is that we allow $\ell(\theta, \theta') = 0$ for $\theta \neq \theta'$.

Minimax posterior concentration rate. As our sequence $(\varepsilon_n)_{n\geqslant 1}$ does not depend on the specific $\theta \in \Theta$ at hand, we may study uniform posterior concentration rates: sequences $(\varepsilon_n)_{n\geqslant 1}$ such that

$$\inf_{\theta\in\Theta} \mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) \leqslant \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \longrightarrow 1 \qquad \text{as } n \to +\infty\,. \tag{6.15}$$

The minimax posterior concentration rate is given by a sequence $(\varepsilon_n)_{n\geqslant 1}$ such that (6.15) holds for some prior π while there exists a constant $\gamma \in (0,1)$ such that for all priors π' on Θ,

$$\limsup_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta',\theta) \leqslant \gamma\,\varepsilon_n \,\big|\, X_{1:n}\big)\Big] < 1\,.$$

We focus on proving the latter statement and provide a general technique to do so. Though we only illustrate it in the finite-dimensional Gaussian setting, adapting it to, e.g., the nonparametric regression problem of Appendix 6.8.1 would add no technical difficulty.

Proposition 6.4.1 (A posterior concentration lower bound in the finite-dimensional Gaussian model)**.**
Let $d \geqslant 1$ be the ambient dimension, $n \geqslant 1$ the sample size, and $\sigma > 0$ the standard deviation. Assume we observe an n–sample $X_{1:n} = (X_1, \dots, X_n)$ distributed according to $\mathcal{N}(\theta, \sigma^2 I_d)$ for some unknown $\theta \in \mathbb{R}^d$. Let π' be any prior distribution on $\mathbb{R}^d$. Then the posterior distribution $\mathbb{P}_{\pi'}(\,\cdot \,|\, X_{1:n})$ defined in (6.14) satisfies, for the Euclidean loss $\ell(\theta',\theta) = \|\theta' - \theta\|_2$ and for $\varepsilon_n = (\sigma/8)\sqrt{d/n}$,

$$\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \|\theta' - \theta\|_2 \leqslant \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \leqslant c_d\,,$$

where $(c_d)_{d\geqslant 1}$ is a decreasing sequence such that $c_1 \leqslant 0.55$, $c_2 \leqslant 0.37$, and $c_d \to 0.21$ as $d \to +\infty$.

This proposition indicates that the best possible posterior concentration rate is at best $\sigma\sqrt{d/n}$ up to a multiplicative constant; actually, this order of magnitude is the best achievable posterior concentration rate, see, e.g., Le Cam and Yang [2000, Chapter 8].

There are at least two ways to prove the lower bound of Proposition 6.4.1. A first one is to use a well-known conversion of "good" Bayesian posteriors into "good" point estimators, which indicates that lower bounds for point estimation can be turned into lower bounds for posterior concentration. For the sake of completeness, we recall this conversion in Appendix 6.8.2 and provide a nonasymptotic variant of Theorem 2.5 by Ghosal et al. [2000].

The second method—followed in the proof below—is however more direct. We use our most general continuous Fano's inequality with the random variables $Z_\theta = \mathbb{P}_{\pi'}\big(\theta' : \|\theta' - \theta\|_2 \leqslant \varepsilon_n \,\big|\, X_{1:n}\big) \in [0,1]$.

Proof. We may assume, with no loss of generality, that the probability space on which $X_{1:n}$ is defined is $(\mathbb{R}^d)^n$ endowed with its Borel σ–field and the probability measure $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^{\otimes n}$. Let ν denote the uniform distribution on the Euclidean ball $B(0, \rho\varepsilon_n) = \{u \in \mathbb{R}^d : \|u\|_2 \leqslant \rho\varepsilon_n\}$ for some $\rho > 1$ to be determined by the analysis. Then, by the continuous Fano inequality in the form given by the combination of (6.9) and (6.13), with $\mathbb{Q}_\theta = \mathbb{P}_0 = \mathcal{N}(0, \sigma^2)^{\otimes n}$, where 0 denotes the null vector of $\mathbb{R}^d$, and with the $[0,1]$–valued random variables $Z_\theta = \mathbb{P}_{\pi'}(\theta' : \|\theta' - \theta\|_2 \leqslant \varepsilon_n \,|\, X_{1:n})$, we have

$$\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\big[Z_\theta\big] \leqslant \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_\theta\big[Z_\theta\big]\mathrm{d}\nu(\theta) \leqslant \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\mathrm{d}\nu(\theta) + \sqrt{\frac{\displaystyle\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0)\mathrm{d}\nu(\theta)}{\displaystyle -\log \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\mathrm{d}\nu(\theta)}}$$
$$\leqslant \left(\frac{1}{\rho}\right)^d + \sqrt{\frac{n\rho^2\varepsilon_n^2/(2\sigma^2)}{d\log\rho}}\,, \tag{6.16}$$

where the last inequality follows from (6.17) and (6.18) below. First note that, by independence, $\mathrm{KL}\big(\mathbb{P}_\theta, \mathbb{P}_0\big) = n\mathrm{KL}\big(\mathcal{N}(\theta,\sigma^2), \mathcal{N}(0,\sigma^2)\big) = n\,\|\theta\|_2^2/(2\sigma^2)$, so that

$$\int_{B(0,\rho\varepsilon_n)} \mathrm{KL}\big(\mathbb{P}_\theta, \mathbb{P}_0\big)\mathrm{d}\nu(\theta) = \frac{n}{2\sigma^2}\int_{B(0,\rho\varepsilon_n)} \|\theta\|_2^2\,\mathrm{d}\nu(\theta) \leqslant \frac{n\rho^2\varepsilon_n^2}{2\sigma^2}\,. \tag{6.17}$$

Second, using the Fubini-Tonelli theorem (twice) and the definition of

$$Z_\theta = \mathbb{P}_{\pi'}(\theta' : \|\theta' - \theta\|_2 \leqslant \varepsilon_n \,|\, X_{1:n}) = \mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\big[\mathbb{1}_{\{\|\theta'-\theta\|_2\leqslant\varepsilon_n\}}\big]\,,$$

we can see that

$$\begin{aligned} q \overset{\text{def}}{=} \int_{B(0,\rho\varepsilon_n)} \mathbb{E}_0\big[Z_\theta\big]\mathrm{d}\nu(\theta) &= \mathbb{E}_0\left[\int_{B(0,\rho\varepsilon_n)} \mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\big[\mathbb{1}_{\{\|\theta'-\theta\|_2\leqslant\varepsilon_n\}}\big]\mathrm{d}\nu(\theta)\right] \\ &= \mathbb{E}_0\left[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\left[\int_{B(0,\rho\varepsilon_n)} \mathbb{1}_{\{\|\theta'-\theta\|_2\leqslant\varepsilon_n\}}\mathrm{d}\nu(\theta)\right]\right] \\ &= \mathbb{E}_0\Big[\mathbb{E}_{\theta'\sim\mathbb{P}_{\pi'}(\cdot\,|\,X_{1:n})}\Big[\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big)\Big]\Big] \leqslant \left(\frac{1}{\rho}\right)^d, \end{aligned} \tag{6.18}$$

where to get the last inequality we used the fact that $\nu\big(B(\theta',\varepsilon_n)\cap B(0,\rho\varepsilon_n)\big)$ is the ratio of the volume of the (possibly truncated) Euclidean ball $B(\theta',\varepsilon_n)$ of radius ε_n and center θ' with the volume of the support of ν, namely, the larger Euclidean ball $B(0,\rho\varepsilon_n)$, in dimension d.

The proof is then concluded by recalling that $\rho > 1$ was a parameter of the analysis and by picking, e.g., $\varepsilon_n = (\sigma/8)\sqrt{d/n}$: by (6.16), we have

$$\inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \|\theta' - \theta\|_2 \leqslant \varepsilon_n \,\big|\, X_{1:n}\big)\Big] = \inf_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\big[Z_\theta\big] \leqslant \inf_{\rho>1}\left\{\left(\frac{1}{\rho}\right)^d + \frac{\rho}{8\sqrt{2\log\rho}}\right\} \overset{\text{def}}{=} c_d\,.$$

We can see that $c_1 \leqslant 0.55$ and $c_2 \leqslant 0.37$ via the respective choices $\rho = 5$ and $\rho = 3$, while the fact that the limit is smaller than (and actually equal to) $\sqrt{e}/8 \leqslant 0.21$ follows from the choice $\rho = \sqrt{e}$.

Note that, when using (6.13) above, we implicitly assumed that the quantity q in (6.18) lies in $(0,1)$. The fact that $q < 1$ follows directly from the upper bound $(1/\rho)^d$ and from $\rho > 1$. Besides, the condition $q > 0$ is met as soon as $\mathbb{P}_0\big(\mathbb{P}_{\pi'}(B(0,\varepsilon_n) \,|\, X_{1:n}) > 0\big) > 0$; indeed, for $\theta' \in B(0,\varepsilon_n)$, we have $\nu\big(B(\theta',\varepsilon_n) \cap B(0,\rho\varepsilon_n)\big) > 0$ and thus q appears in the last equality of (6.18) as being lower bounded by the expectation of a positive function over a set with positive probability. If on the contrary $\mathbb{P}_0\big(\mathbb{P}_{\pi'}(B(0,\varepsilon_n) \,|\, X_{1:n}) > 0\big) = 0$, then $\mathbb{P}_0(Z_0 > 0) = 0$, so that $\inf_\theta \mathbb{E}_\theta[Z_\theta] = \mathbb{E}_0[Z_0] = 0$, which immediately implies the bound of Proposition 6.4.1. □

Remark 6.4.2. *Though the lower bound of Proposition 6.4.1 is only stated for the posterior distributions $\mathbb{P}_{\pi'}(\,\cdot \,|\, X_{1:n})$, it is actually valid for any transition kernel $Q(\,\cdot \,|\, X_{1:n})$. This is because the proof above relies on general information-theoretic arguments and does not use the particular form of $\mathbb{P}_{\pi'}(\,\cdot \,|\, X_{1:n})$. This is in the same spirit as for minimax lower bounds for point estimation.*

In Section 6.5.2 we derive another type of posterior concentration lower bound that is no longer uniform. More precisely, we prove a distribution-dependent lower bound that specifies how the posterior mass fails to concentrate on ε_n–neighborhoods of θ for every $\theta \in \Theta$.

6.4.2 Lower bounds in robust sequential learning with sparse losses

We consider a framework of robust sequential learning called prediction of individual sequences. Its origins and core results are described in the monography by Cesa-Bianchi and Lugosi [2006]. In its simplest version, a decision-maker and an environment play repeatedly as follows: at each round $t \geqslant 1$, and simultaneously, the environment chooses a vector of losses $\ell_t = (\ell_{1,t}, \dots, \ell_{N,t}) \in [0,1]^N$ while the decision-maker picks an index $I_t \in \{1, \dots, N\}$, possibly at random. Both players then observe ℓ_t and I_t. The decision-maker wants to minimize her cumulative regret, the difference between her cumulative loss and the cumulative loss associated with the best constant choice of an index: for $T \geqslant 1$,

$$R_T = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{k=1,\dots,N} \sum_{t=1}^{T} \ell_{k,t} \,.$$

In this setting the optimal regret in the worst-case is of the order of $\sqrt{T \log(N)}$. Cesa-Bianchi et al. [1997] exhibited an asymptotic lower bound of $\sqrt{T \log(N)/2}$, based on the central limit theorem and on the fact that the expectation of the maximum of N independent standard Gaussian random variables is of the order of $\sqrt{\log(N)}$. To do so, they considered stochastic environments drawing independently the loss vectors ℓ_t according to a well-chosen distribution.

Cesa-Bianchi et al. [2005] extended this result to a variant called label-efficient prediction, in which loss vectors are observed upon choosing and with a budget constraint: no

more than m observations within T rounds. They prove an optimal and non-asymptotic lower bound on the regret of the order of $T\sqrt{\log(N)/m}$, based on several applications of Fano's inequality to deterministic strategies of the decision-maker, and then, an application of Fubini's theorem to handle general, randomized, strategies. Our re-shuffled proof technique below shows that a single application of Fano's inequality to general strategies would be sufficient there (details omitted).

Recently, Kwon and Perchet [2016] considered a setting of sparse loss vectors, in which at each round at most s of the N components of the loss vectors ℓ_t are different from zero. They prove an optimal and asymptotic lower bound on the regret of the order of $\sqrt{Ts\log(N)/N}$, which generalizes the result for the basic framework, in which $s = N$. Their proof is an extension of the proof of Cesa-Bianchi et al. [1997] and is based on the central limit theorem together with additional technicalities, e.g., the use of Slepian's lemma to deal with some dependencies arising from the sparsity assumption.

The aim of this section is to provide a short and elementary proof of this optimal $\sqrt{Ts\log(N)/N}$ bound. As a side result, our bound will even be non-asymptotic. The expectation in the statement below is with respect to the internal randomization used by the decision-maker's strategy.

Theorem 6.4.3. *For all strategies of the decision-maker, for all $N \geqslant 2$ and all $T > N\log(N)/(16s)$, there exists a fixed-in-advance sequence of loss vectors $\ell_1, \ldots, \ell_T$ in $[0,1]^N$ that are each s–sparse such that*

$$\mathbb{E}\big[R_T\big] = \sum_{t=1}^{T} \mathbb{E}\big[\ell_{I_t,t}\big] - \min_{k=1,\ldots,N} \sum_{t=1}^{T} \ell_{k,t} \geqslant \frac{1}{32}\sqrt{T\frac{s}{N}\log N}\,.$$

Proof. We fix $\varepsilon \in \big(0,\, s/(2N)\big)$ and consider, as Kwon and Perchet [2016] did, independent and identically distributed loss vectors $\ell_t \in [0,1]^N$, drawn according to one distribution among P_i, where $1 \leqslant i \leqslant N$. Each distribution P_i over $[0,1]^N$ is defined as the law of a random vector L drawn in two steps as follows. We pick s components uniformly at random among $\{1,\ldots,N\}$. Then, the components k not picked are associated with zero losses, $L_k = 0$. The losses L_k for picked components $k \neq i$ are drawn according to a Bernoulli distribution with parameter $1/2$. If component i is picked, its loss L_i is drawn according to a Bernoulli distribution with parameter $1/2 - \varepsilon N/s$. The loss vector $L \in [0,1]^N$ thus generated is indeed s–sparse. We denote by P_i^T the T–th product distribution $P_i \otimes \cdots \otimes P_i$. We will actually identify the underlying probability and the law P_i^T. Finally, we denote the expectation under P_i^T by $\mathbb{E}_i$.

Now, under P_i^T, the components $\ell_{k,t}$ of the loss vectors are all distributed according to Bernoulli distributions, with parameters $s/(2N)$ if $k \neq i$ and $s/(2N) - \varepsilon$ if $k = i$. The expected regret, where the expectation $\mathbb{E}$ is with respect to the strategy's internal randomization and the expectation $\mathbb{E}_i$ is with respect to the random choice of the loss

vectors, is thus larger than

$$\begin{aligned}\mathbb{E}_i\Big[\mathbb{E}\big[R_T\big]\Big] &\geqslant \sum_{t=1}^{T} \mathbb{E}_i\Big[\mathbb{E}\big[\ell_{I_t,t}\big]\Big] - \min_{k=1,\dots,N} \sum_{t=1}^{T} \mathbb{E}_i\big[\ell_{k,t}\big] \\ &= \sum_{t=1}^{T} \frac{s}{2N}\bigg(1 - \varepsilon\,\mathbb{E}_i\Big[\mathbb{E}\big[\mathbb{1}\{I_t = i\}\big]\Big]\bigg) - T\bigg(\frac{s}{2N} - \varepsilon\bigg) \\ &= T\varepsilon\bigg(1 - \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big]\bigg)\,,\end{aligned} \tag{6.19}$$

where

$$F_i(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}\,.$$

All in all, we copied almost word for word the (standard) beginning of the proof by Kwon and Perchet [2016], whose first lower bound is exactly

$$\sup_{\ell_1,\dots,\ell_t} \mathbb{E}\big[R_T\big] \geqslant \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[R_T\big]\Big] \geqslant T\varepsilon\bigg(1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big]\bigg). \tag{6.20}$$

The main differences arise now: we replace a long asymptotic argument (based on the central limit theorem and the study of the limit via Slepian's lemma) by a single application of Fano's inequality.

We introduce the distribution Q over $[0,1]^N$ corresponding to the same randomization scheme as for the P_i, except that no picked component is favored and that all their corresponding losses are drawn according to the Bernoulli distribution with parameter $1/2$. We also denote by $\mathbb{P}$ the probability distribution that underlies the internal randomization of the strategy. An application of Lemma 6.3.2 with $\mathbb{P}_i = \mathbb{P}\otimes P_i^T$ and $\mathbb{Q}_i = \mathbb{P}\otimes Q^T$, using that $F_1(T) + \ldots + F_N(T) = 1$ and thus $(1/N)\sum_{i=1}^{N} \mathbb{E}_Q\big[\mathbb{E}[F_i(T)]\big] = 1/N$, yields

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\Big[\mathbb{E}\big[F_i(T)\big]\Big] \leqslant \frac{1}{N} + \sqrt{\frac{1}{N\log(N)}\sum_{i=1}^{N} \mathrm{KL}\big(\mathbb{P}\otimes P_i^T,\, \mathbb{P}\otimes Q^T\big)}\,. \tag{6.21}$$

By independence, we get, for all i,

$$\mathrm{KL}\big(\mathbb{P}\otimes P_i^T,\, \mathbb{P}\otimes Q^T\big) = \mathrm{KL}\big(P_i^T,\, Q^T\big) = T\,\mathrm{KL}\big(P_i,\, Q\big)\,. \tag{6.22}$$

We now show that

$$\mathrm{KL}\big(P_i,\, Q\big) \leqslant \frac{s}{N}\,\mathrm{kl}\bigg(\frac{1}{2} - \varepsilon\frac{N}{s},\, \frac{1}{2}\bigg)\,. \tag{6.23}$$

Indeed, both P_i and Q can be seen as uniform convex combinations of probability distributions of the following form, indexed by the subsets of $\{1,\dots,N\}$ with s elements

and up to permutations of the Bernoulli distributions in the products below (which does not change the value of the Kullback-Leibler divergences between them):

$$\binom{N-1}{s-1} \text{ distributions of the form (when } i \text{ is picked)}$$
$$\mathrm{Ber}\Big(\frac{1}{2}-\varepsilon\frac{N}{s}\Big)\otimes\bigotimes_{k=2}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0 \qquad \text{and} \qquad \bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0\,,$$

where δ_0 denotes the Dirac mass at 0, and

$$\binom{N-1}{s} \text{ distributions of the form (when } i \text{ is not picked)}$$
$$\bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0 \qquad \text{and} \qquad \bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0\,.$$

Only the first set of distributions contributes to the Kullback-Leibler divergence. By convexity of the Kullback-Leibler divergence (Corollary 6.2.3), we thus get the inequality

$$\begin{aligned}\mathrm{KL}\big(P_i,\,Q\big) &\leqslant \frac{\binom{N-1}{s-1}}{\binom{N}{s}}\mathrm{KL}\Bigg(\mathrm{Ber}\Big(\frac{1}{2}-\varepsilon\frac{N}{s}\Big)\otimes\bigotimes_{k=2}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0,\ \bigotimes_{k=1}^{s}\mathrm{Ber}\Big(\frac{1}{2}\Big)\otimes\bigotimes_{k=s+1}^{N}\delta_0\Bigg)\\ &= \frac{s}{N}\,\mathrm{kl}\Big(\frac{1}{2}-\varepsilon\frac{N}{s},\,\frac{1}{2}\Big)\,,\end{aligned}$$

where the last equality is again by independence. Finally, the lemma stated right after this proof shows that

$$\mathrm{kl}\Big(\frac{1}{2}-\varepsilon\frac{N}{s},\,\frac{1}{2}\Big)\leqslant\frac{4N^2\varepsilon^2}{s^2}\,. \tag{6.24}$$

Combining (6.20)–(6.24), we proved

$$\sup_{\ell_1,\dots,\ell_t}\mathbb{E}\big[R_T\big]\geqslant T\varepsilon\Bigg(1-\frac{1}{N}-\sqrt{\frac{4NT\varepsilon^2}{s\log(N)}}\Bigg)\geqslant T\varepsilon\Big(\frac{1}{2}-c\,\varepsilon\Big)\,,$$

where we used $1/N\leqslant 1/2$ and denoted $c=2\sqrt{NT}/\sqrt{s\log(N)}$. A standard optimization suggest the choice $\varepsilon=1/(4c)$, which is valid, i.e., is indeed $<s/(2N)$ as required, as soon as $T>N\log(N)/(16s)$. We get a lower bound $T\varepsilon/4$, which is the claimed bound. □

Lemma 6.4.4. *For all $p\in(0,1)$, for all $\varepsilon\in(0,p)$,*

$$\mathrm{kl}(p-\varepsilon,p)\leqslant\frac{\varepsilon^2}{p(1-p)}\,.$$

Proof. This result is a special case of the fact that the KL divergence is upper bounded by the χ^2–divergence. We recall, in our particular case, how this is seen:

$$\begin{aligned}\mathrm{kl}(p-\varepsilon,p) &= (p-\varepsilon)\log\left(1-\frac{\varepsilon}{p}\right)+(1-p+\varepsilon)\log\left(1+\frac{\varepsilon}{1-p}\right)\\ &\leqslant (p-\varepsilon)\frac{-\varepsilon}{p}+(1-p+\varepsilon)\frac{\varepsilon}{1-p}=\frac{\varepsilon^2}{p}+\frac{\varepsilon^2}{1-p}\,,\end{aligned}$$

where we used $\log(1+u)\leqslant u$ for all $u>-1$ to get the stated inequality. □

6.5 Other applications, with $N=1$ pair of distributions

Interestingly, Proposition 6.2.1 can be useful even for $N=1$ pair of distributions. Rewriting it slightly differently, we indeed have, for all distributions $\mathbb{P},\mathbb{Q}$ and all events A,

$$\mathbb{P}(A)\log\left(\frac{1}{\mathbb{Q}(A)}\right)\leqslant \mathrm{KL}(\mathbb{P},\mathbb{Q})+\log(2)\,.$$

Solving for $\mathbb{Q}(A)$—and not for $\mathbb{P}(A)$ as was previously the case—we get

$$\mathbb{Q}(A)\geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})+\log(2)}{\mathbb{P}(A)}\right)\,, \tag{6.25}$$

where the above inequality is true even if $\mathbb{P}(A)=0$ or $\mathrm{KL}(\mathbb{P},\mathbb{Q})=+\infty$. More generally, for all distributions $\mathbb{P},\mathbb{Q}$ and all $[0,1]$–valued random variables Z, we have, by Corollary 6.3.1,

$$\mathbb{E}_{\mathbb{Q}}[Z]\geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})+\log(2)}{\mathbb{E}_{\mathbb{P}}[Z]}\right)\,, \tag{6.26}$$

where again the above inequality is true even if $\mathbb{E}_{\mathbb{P}}[Z]=0$ or $\mathrm{KL}(\mathbb{P},\mathbb{Q})=+\infty$.

The bound (6.25) is similar in spirit to (a consequence of) the Bretagnolle-Huber inequality, recalled and actually improved in Section 6.7.3; see details therein, and in particular its consequence (6.44). Both bounds can indeed be useful when $\mathrm{KL}(\mathbb{P},\mathbb{Q})$ is larger than a constant and $\mathbb{P}(A)$ is close to 1.

Next we show two applications of (6.25) and (6.26): a simple proof of a large deviation lower bound for Bernoulli distributions, and a distribution-dependent posterior concentration lower bound.

6.5.1 A simple proof of Cramér's theorem for Bernoulli distributions

The next proposition is a well-known large deviation result on the sample mean of independent and identically distributed Bernoulli random variables. It is a particular case of Cramér's theorem that dates back to Cramér [1938], Chernoff [1952]; see also Cerf and Petit [2011] for further references and a proof in a very general context. Thanks to Fano's inequality (6.25), the proof of the lower bound that we provide below avoids any explicit change of measure (see the remark after the proof).

Proposition 6.5.1 (Cramér's theorem for Bernoulli distributions)**.** *Let $\theta \in (0,1)$. Assume that $X_1, \dots, X_n$ are independent and identically distributed random variables drawn from* $\mathrm{Ber}(\theta)$*. Denoting by $\mathbb{P}_\theta$ the underlying probability measure, we have, for all $x \in (\theta, 1)$,*

$$\lim_{n\to+\infty} \frac{1}{n} \log \mathbb{P}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} X_i > x\right) = -\mathrm{kl}(x,\theta)\,.$$

Proof. We set $\overline{X}_n \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n X_i$. For the convenience of the reader we first briefly recall how to prove the upper bound, and then proceed with a new proof for the lower bound.

Upper bound: By the Cramér-Chernoff method and the duality formula for the Kullback-Leibler divergence between Bernoulli distributions (see, e.g., Boucheron et al. 2013, pages 21–24), we have, for all $n \geqslant 1$,

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \leqslant \exp\left(-n \sup_{\lambda>0}\Big\{\lambda x - \log \mathbb{E}_\theta\big[e^{\lambda X_1}\big]\Big\}\right) = \exp\Big(-n\mathrm{kl}(x,\theta)\Big)\,, \tag{6.27}$$

that is,

$$\forall n \geqslant 1, \qquad \frac{1}{n}\log \mathbb{P}_\theta\big(\overline{X}_n > x\big) \leqslant -\mathrm{kl}(x,\theta)\,.$$

Lower bound: Choose $\varepsilon > 0$ small enough such that $x+\varepsilon < 1$. As in the proof of Proposition 6.8.1, we may assume with no loss of generality that the underlying distribution is $\mathbb{P}_\theta = \mathrm{Ber}(\theta)^{\otimes n}$. By Fano's inequality in the form (6.25) with the distributions $\mathbb{P} = \mathbb{P}_{x+\varepsilon}$ and $\mathbb{Q} = \mathbb{P}_\theta$, and the event $A = \big\{\overline{X}_n > x\big\}$, we have

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant \exp\left(-\frac{\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) + \log(2)}{\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big)}\right)\,.$$

Noting that $\mathrm{KL}(\mathbb{P}_{x+\varepsilon}, \mathbb{P}_\theta) = n\mathrm{kl}(x+\varepsilon,\theta)$ we get

$$\mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant \exp\left(-\frac{n\mathrm{kl}(x+\varepsilon,\theta) + \log 2}{\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big)}\right) \geqslant \exp\left(-\frac{n\mathrm{kl}(x+\varepsilon,\theta) + \log 2}{1 - e^{-n\mathrm{kl}(x,x+\varepsilon)}}\right)\,, \tag{6.28}$$

where the last bound follows from $\mathbb{P}_{x+\varepsilon}\big(\overline{X}_n > x\big) = 1 - \mathbb{P}_{x+\varepsilon}\big(\overline{X}_n \leqslant x\big) \geqslant 1 - e^{-n\mathrm{kl}(x,x+\varepsilon)}$ by a derivation similar to (6.27) above. Taking the logarithms of both sides and letting $n \to +\infty$ finally yields

$$\liminf_{n\to+\infty} \frac{1}{n}\log \mathbb{P}_\theta\big(\overline{X}_n > x\big) \geqslant -\mathrm{kl}(x+\varepsilon,\theta)\,.$$

We conclude the proof by letting $\varepsilon \to 0$, and by combining the upper and lower bounds. □

Comparison with an historical proof. A classical proof for the lower bound relies on the same change of measure as the one used above, i.e., that transports the measure $\mathrm{Ber}(\theta)^{\otimes n}$ to $\mathrm{Ber}(x+\varepsilon)^{\otimes n}$. The bound (6.27), or any other large deviation inequality, is also typically used therein. However, the change of measure is usually carried out explicitly by writing

$$\begin{aligned}\mathbb{P}_\theta\big(\overline{X}_n > x\big) &= \mathbb{E}_\theta\Big[\mathbb{1}_{\{\overline{X}_n > x\}}\Big] \\ &= \mathbb{E}_{x+\varepsilon}\Big[\mathbb{1}_{\{\overline{X}_n > x\}}\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mathbb{P}_{x+\varepsilon}}(X_1,\dots,X_n)\Big] = \mathbb{E}_{x+\varepsilon}\Big[\mathbb{1}_{\{\overline{X}_n > x\}}e^{-n\,\widehat{\mathrm{KL}}_n}\Big]\,,\end{aligned}$$

where the empirical Kullback-Leibler divergence $\widehat{\mathrm{KL}}_n$ is defined by

$$\begin{aligned}\widehat{\mathrm{KL}}_n &\stackrel{\mathrm{def}}{=} \frac{1}{n}\log\bigg(\frac{\mathrm{d}\mathbb{P}_{x+\varepsilon}}{\mathrm{d}\mathbb{P}_\theta}(X_1,\dots,X_n)\bigg) \\ &= \frac{1}{n}\sum_{i=1}^n\bigg(\mathbb{1}_{\{X_i=1\}}\log\bigg(\frac{x+\varepsilon}{\theta}\bigg) + \mathbb{1}_{\{X_i=0\}}\log\bigg(\frac{1-(x+\varepsilon)}{1-\theta}\bigg)\bigg)\,.\end{aligned}$$

The empirical Kullback-Leibler divergence $\widehat{\mathrm{KL}}_n$ is then compared to its limit $\mathrm{kl}(x+\varepsilon,\theta)$ via the law of large numbers. On the contrary, our short proof above bypasses any call to the law of large numbers and does not perform the change of measure explicitely, in the same spirit as for the bandit lower bounds derived by Kaufmann et al. [2016] and in Chapter 2. Note that the different and more general proof of Cerf and Petit [2011] also bypassed any call to the law of large numbers thanks to other convex duality arguments.

6.5.2 Distribution-dependent posterior concentration lower bounds

In this section we consider the same Bayesian setting as the one described at the beginning of Section 6.4.1. In addition, we define the global modulus of continuity between KL and ℓ around $\theta \in \Theta$ and at scale $\varepsilon_n > 0$ by

$$\psi\big(\varepsilon_n,\theta,\ell\big) \stackrel{\mathrm{def}}{=} \inf\Big\{\mathrm{KL}\big(P_{\theta'},P_\theta\big)\,:\,\ell(\theta',\theta) \geqslant 2\varepsilon_n,\ \theta' \in \Theta\Big\}\,;$$

the infimum is set to $+\infty$ if the set is empty.

Next we provide a distribution-dependent lower bound for posterior concentration rates, that is, a lower bound that holds true for every $\theta \in \Theta$, as opposed to the minimax lower bound of Section 6.4.1. Note however that we are here in a slightly different regime than in Section 6.4.1, where we addressed cases for which the uniform posterior concentration condition (6.30) below was proved to be impossible at scale ε_n (and actually took place at a slightly larger scale ε'_n).

Theorem 6.5.2 (Distribution-dependent posterior concentration lower bound)**.** *Assume that the posterior distribution $\mathbb{P}_\pi(\,\cdot\,|\,X_{1:n})$ satisfies the uniform concentration condition*

$$\inf_{\theta\in\Theta}\ \mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \longrightarrow 1 \qquad \text{as } n \to +\infty\,.$$

Then, for all $\theta \in \Theta$ and $c > 1$, for all n large enough,

$$\mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) > \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant 2^{-c} \exp\Big(-c\, n\, \psi(\varepsilon_n, \theta, \ell)\Big)\,. \tag{6.29}$$

The conclusion can be stated equivalently as: for all $\theta \in \Theta$,

$$\liminf_{n \to +\infty} \frac{\log\Big(\mathbb{E}_\theta\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) > \varepsilon_n \big| X_{1:n}\big)\Big]\Big)}{\log(2) + n\, \psi(\varepsilon_n, \theta, \ell)} \geqslant -1\,.$$

The above theorem is greatly inspired from Theorem 2.1 by Hoffmann et al. [2015]. Our Fano's inequality (6.26) however makes the proof more direct: the change-of-measure carried out by Hoffmann et al. [2015] is now implicit, and no proof by contradiction is required. We also bypass one technical assumption (see the discussion after the proof).

Proof. We fix $\theta \in \Theta$ and $c > 1$. By the uniform concentration condition, there exists $n_0 \geqslant 1$ such that, for all $n \geqslant n_0$,

$$\inf_{\theta^\star \in \Theta} \mathbb{E}_{\theta^\star}\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta^\star) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant \frac{1}{c}\,. \tag{6.30}$$

We now fix $n \geqslant n_0$ and consider any $\theta^\star \in \Theta$ such that $\ell\big(\theta^\star, \theta\big) \geqslant 2\varepsilon_n$. Using Fano's inequality in the form of (6.26) with the distributions $\mathbb{P} = P_{\theta^\star}^{\otimes n}$ and $\mathbb{Q} = P_\theta^{\otimes n}$, together with the $[0,1]$–valued random variable $Z_\theta = \mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta) > \varepsilon_n \,\big|\, X_{1:n}\big)$, we get

$$\mathbb{E}_\theta\big[Z_\theta\big] \geqslant \exp\left(-\frac{\mathrm{KL}\Big(P_{\theta^\star}^{\otimes n}, P_\theta^{\otimes n}\Big) + \log 2}{\mathbb{E}_{\theta^\star}\big[Z_\theta\big]}\right) = \exp\left(-\frac{n\mathrm{KL}\big(P_{\theta^\star}, P_\theta\big) + \log 2}{\mathbb{E}_{\theta^\star}\big[Z_\theta\big]}\right). \tag{6.31}$$

By the triangle inequality and the assumption $\ell\big(\theta^\star, \theta\big) \geqslant 2\varepsilon_n$ we can see that $\big\{\theta' : \ell(\theta', \theta) > \varepsilon_n\big\} \supseteq \big\{\theta' : \ell(\theta', \theta^\star) < \varepsilon_n\big\}$, so that

$$\mathbb{E}_{\theta^\star}\big[Z_\theta\big] \geqslant \mathbb{E}_{\theta^\star}\Big[\mathbb{P}_\pi\big(\theta' : \ell(\theta', \theta^\star) < \varepsilon_n \big| X_{1:n}\big)\Big] \geqslant \frac{1}{c}$$

by the uniform lower bound (6.30). Substituting the above inequality into (6.31) then yields

$$\mathbb{E}_\theta\big[Z_\theta\big] \geqslant \exp\Big(-c\Big(n\mathrm{KL}\big(P_{\theta^\star}, P_\theta\big) + \log 2\Big)\Big)\,.$$

To conclude the proof, it suffices to take the supremum of the right-hand side over all $\theta^\star \in \Theta$ such that $\ell\big(\theta^\star, \theta\big) \geqslant 2\varepsilon_n$, and to identify the definition of $\psi\big(\varepsilon_n, \theta, \ell\big)$. □

Note that, at first sight, our result may seem a little weaker than Hoffmann et al. [2015, Theorem 2.1], because we only define $\psi\big(\varepsilon_n, \theta, \ell\big)$ in terms of KL instead of a general pre-metric d: in other words, we only consider the case $d(\theta, \theta') = \sqrt{\mathrm{KL}(P_{\theta'}, P_\theta)}$.

However, it is still possible to derive a bound in terms of an arbitrary pre-metric d by comparing d and KL after applying Theorem 6.5.2.

In the case of the pre-metric $d(\theta,\theta') = \sqrt{\mathrm{KL}(P_{\theta'}, P_\theta)}$, we bypass an additional technical assumption used for the the similar lower bound of Hoffmann et al. [2015, Theorem 2.1]; namely, that there exists a constant $C > 0$ such that

$$\sup_{\theta,\theta'} P_{\theta'}^{\otimes n}\Big(\mathcal{L}_n(\theta') - \mathcal{L}_n(\theta) \geqslant Cn\mathrm{KL}\big(P_{\theta'}, P_\theta\big)\Big) \longrightarrow 0 \qquad \text{as } n \to +\infty\,,$$

where the supremum is over all $\theta, \theta' \in \Theta$ satisfying $\psi\big(\varepsilon_n, \theta, \ell\big) \leqslant \mathrm{KL}\big(P_{\theta'}, P_\theta\big) \leqslant 2\psi\big(\varepsilon_n, \theta, \ell\big)$, and where $\mathcal{L}_n(\theta) = \sum_{i=1}^n \log\big(\mathrm{d}P_\theta/\mathrm{d}\mathfrak{m}\big)(X_i)$ denotes the log-likelihood function with respect to a common dominating measure $\mathfrak{m}$. Besides, we get an improved constant in the exponential in (6.29), with respect to Hoffmann et al. [2015, Theorem 2.1]: by a factor of $3C/c$, which, since $C \geqslant 1$ in most cases, is $3C/c \approx 3C \geqslant 3$ when $c \approx 1$. (A closer look at their proof can yield a constant arbitrarily close to $2C$, which is still larger than our c by a factor of $2C/c \approx 2C \geqslant 2$.)

6.6 References and comparison to the literature

We discuss in this section how novel (or not novel) our results and approaches are.

Main innovations. We could find no reference indicating that the alternative distributions $\mathbb{Q}_i$ and $\mathbb{Q}_\theta$ could vary and do not need to be set to a fixed alternative $\mathbb{Q}_0$, nor that arbitrary $[0,1]$–valued random variables Z_i or Z_θ could be considered. In particular, to the best of our knowledge, reduction (6.9) is a new result. We provide two novel applications with $[0,1]$–valued random variables in Section 6.4.

Also, as we discuss below in detail when referring to the work of Birgé [2005], results like Lemma 6.3.2 provide an interpolation between the most classical versions of Fano's inequality with a $\log(2)$ factor and Pinsker's inequality. Typically, depending on $N \geqslant 3$ or $N = 2$, one or the other lemma had to be used, while Lemma 6.3.2 can be used in all cases.

What on the contrary was already known. The inequalities (6.10) are folklore knowledge. The first inequality in (6.11) can be found in Guntuboyina [2011]; the second inequality is a new (immediate) consequence. The first inequality in (6.13) is a consequence, which we derived on our own, of a refined Pinsker's inequality stated by Ordentlich and Weinberger [2005], while the second inequality is ours again.

Reduction (6.9) is new, as we indicated, but all other reductions were known, though sometimes proved in a more involved way. Reduction (6.2) and (6.6) were already known and used by Han and Verdú [1994, Theorems 2, 7 and 8]. Reduction (6.7) is stated in spirit by Chen et al. [2016] with a constant alternative $\mathbb{Q}_\theta \equiv \mathbb{Q}$; see also a detailed discussion and comparison below between their approach and the general approach we took in Section 6.3. We should also mention that Duchi and Wainwright [2013] provided

preliminary (though more involved) results towards the continuous reduction (6.7). Reduction (6.8) is stated in a special case in Gushchin [2003], where $Z_1 + \ldots + Z_N = 1$. We also note that while we only discussed Kullback-Leibler divergences so far, all reductions (6.2) and (6.6)–(6.9) extend to f–divergences, as noted already by Gushchin [2003], see also Chen et al. [2016]. We state this extension to f–divergences in Section 6.8.5 of the appendix.

That the sets A_i considered in the reductions (6.2) and (6.6) form a partition of the underlying measurable space or that the random variables Z_i sum up to 1 in (6.8) were typical requirements in the literature until recently. Chen et al. [2016] noted in spirit that the requirement of forming a partition was unnecessary, which we too had been aware of as early as Stoltz [2007], where we also already mentioned the fact that in particular the alternative distribution $\mathbb{Q}$ had not to be fixed and could depend on i or θ.

Finally, the conjunction of a reduction (6.2) or (6.6)–(6.9) and a lower bound on the kl function was already present in Han and Verdú [1994]. Other, more information-theoretic statements and proof techniques of Fano's inequalities for finitely many hypotheses as in Proposition 6.2.1 can be found, e.g., in Cover and Thomas [2006, Theorem 2.11.1], Yu [1997, Lemma 3] or Ibragimov and Has'minskii [1981, Chapter VII, Lemma 1.1] (they resort to classical formulas on the Shannon entropy, the conditional entropy, and the mutual information).

6.6.1 On the "generalized Fanos's inequality" of Chen et al. [2016]

The Bayesian setting considered is the following; it generalizes the setting of Han and Verdú [1994], whose results we discuss in a remark after the proof of Proposition 6.6.1.

A parameter space $(\Theta, \mathcal{G})$ is equipped with a prior probability measure ν. A family of probability distributions $(\mathbb{P}_\theta)_{\theta\in\Theta}$ over a measurable space $(\Omega, \mathcal{F})$, some outcome space $(\mathcal{X}, \mathcal{E})$, e.g., $\mathcal{X} = \mathbb{R}^n$, and a random variable $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{E})$ are considered. We denote by $\mathbb{E}_\theta$ the expectation under $\mathbb{P}_\theta$. Of course we may have $(\Omega, \mathcal{F}) = (\mathcal{X}, \mathcal{E})$ and X be the identity, in which case $\mathbb{P}_\theta$ will be the law of X under $\mathbb{P}_\theta$.

The goal is either to estimate θ or to take good actions: we consider a measurable target space $(\mathcal{A}, \mathcal{H})$, that may or may not be equal to Θ. The quality of a prediction or of an action is measured by a measurable loss function $L : \Theta \times \mathcal{A} \to [0, 1]$. The random variable X is our observation, based on which we construct a $\sigma(X)$–measurable random variable $\widehat{a}$ with values in $\mathcal{A}$. Putting as side all measurability issues (here and in the rest of this subsection), the risk of $\widehat{a}$ in this model equals

$$R(\widehat{a}) = \int_\Theta \mathbb{E}_\theta\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta)$$

and the Bayes risk in this model is the smallest such possible risk,

$$R_{\text{Bayes}} = \inf_{\widehat{a}} R(\widehat{a}) \,,$$

where the infimum is over all $\sigma(X)$–measurable random variables with values in $\mathcal{A}$.

Chen et al. [2016] call their main result (Corollary 5) a "generalized Fano's inequality;" we state it below not only for $\{0,1\}$–valued loss functions L as in the original article but for any $[0,1]$–valued loss functions, as we are able to prove it for any such loss function. The reason behind this extension is that we not only have the reduction (6.7) with events, but we also have the reduction (6.9) with $[0,1]$–valued random variables. We also feel that our proof technique is more direct and more natural.

We only deal with with Kullback-Leibler divergences, but the result and proof below readily extend to f–divergences.

Proposition 6.6.1. *In the setting described above, the Bayes risk is always larger than*

$$R_{\text{Bayes}} \geqslant 1 + \frac{\left(\inf\limits_{\mathbb{Q}} \int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}) \mathrm{d}\nu(\theta)\right) + \log\left(1 + \inf\limits_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \mathrm{d}\nu(\theta)\right)}{\log\left(1 - \inf\limits_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \mathrm{d}\nu(\theta)\right)},$$

where the infimum in the numerator is over all probability measures $\mathbb{Q}$ over $(\Omega, \mathcal{F})$.

Proof. We fix $\widehat{a}$ and an alternative $\mathbb{Q}$. The combination of (6.9) and (6.11), with $Z_\theta = 1 - L\big(\theta, \widehat{a}\big)$, yields

$$1 - \int_\Theta \mathbb{E}_\theta\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta) \leqslant \frac{\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}) \mathrm{d}\nu(\theta) + \log\big(2 - q_{\widehat{a}}\big)}{\log\big(1/q_{\widehat{a}}\big)}, \tag{6.32}$$

where $\mathbb{E}_\mathbb{Q}$ denotes the expectation with respect to $\mathbb{Q}$ and

$$q_{\widehat{a}} = 1 - \int_\Theta \mathbb{E}_\mathbb{Q}\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta)\,.$$

As $q \mapsto 1/\log(1/q)$ and $q \mapsto \log(2-q)/\log(1/q)$ are both increasing, taking the supremum over the $\sigma(X)$–measurable random variables $\widehat{a}$ in both sides of (6.32) gives

$$1 - R_{\text{Bayes}} \leqslant \frac{\int_\Theta \mathrm{KL}(\mathbb{P}_\theta, \mathbb{Q}) \mathrm{d}\nu(\theta) + \log\big(2 - q^\star\big)}{\log\big(1/q^\star\big)} \tag{6.33}$$

where

$$q^\star = \sup_{\widehat{a}} q_{\widehat{a}} = 1 - \inf_{\widehat{a}} \int_\Theta \mathbb{E}_\mathbb{Q}\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta) = 1 - \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \mathrm{d}\nu(\theta)\,, \tag{6.34}$$

as is proved below. Taking the infimum of the right-hand side of (6.33) over $\mathbb{Q}$ and rearranging concludes the proof.

It only remains to prove the last inequality of (6.34) and actually, as constant elements $a \in \mathcal{A}$ are special cases of random variables $\widehat{a}$, we only need to prove that

$$\inf_{\widehat{a}} \int_\Theta \mathbb{E}_\mathbb{Q}\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta) \geqslant \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \mathrm{d}\nu(\theta)\,. \tag{6.35}$$

Now, each $\widehat{a}$ that is $\sigma(X)$–measurable can be rewritten $\widehat{a} = \overline{a}(X)$ for some measurable function $\overline{a} : \mathcal{X} \to \mathcal{A}$; then, by the Fubini-Tonelli theorem:

$$\begin{aligned}\int_\Theta \mathbb{E}_\mathbb{Q}\Big[L\big(\theta, \widehat{a}\big)\Big] \mathrm{d}\nu(\theta) &= \int_\mathcal{X} \left(\int_\Theta L\big(\theta, \overline{a}(x)\big) \mathrm{d}\nu(\theta) \right) \mathrm{d}\mathbb{Q}(x) \\ &\geqslant \int_\mathcal{X} \left(\inf_{a \in \mathcal{A}} \int_\Theta L\big(\theta, a\big) \mathrm{d}\nu(\theta) \right) \mathrm{d}\mathbb{Q}(x)\,,\end{aligned}$$

which proves (6.35). □

Remark 6.6.2. *As mentioned by Chen et al. [2016], one of the major results of Han and Verdú [1994], namely, their Theorem 8, is a special case of Proposition 6.6.1, with* $\Theta = \mathcal{A}$ *and the loss function* $L(\theta, \theta') = \mathbb{1}\{\theta \neq \theta'\}$. *The (opposite of the) denominator in the lower bound on the Bayes risk then takes the simple form*

$$-\log\left(1 - \inf_{\theta' \in \Theta} \int_\Theta L(\theta, \theta') \mathrm{d}\nu(\theta)\right) = \log\left(\sup_{\theta \in \Theta} \nu\big(\{\theta\}\big)\right) \overset{\text{def}}{=} H_\infty(\nu)\,,$$

which is called the infinite-order Rényi entropy of the probability distribution ν. *Han and Verdú [1994] only dealt with the case of discrete sets* Θ *but the extension to continuous* Θ *is immediate, as we showed in Section 6.3.*

6.6.2 Comparison to Birgé [2005]

This version of Fano's inequality is extremely popular among statisticians. We state here a slightly simplified version of the main result by Birgé [2005] (his Corollary 1), inspired by a previous (looser) simplification by Massart [2007]: in Appendix 6.8.3 these two alternative statements are stated, proved, and compared to Theorem 6.6.3. In contrast, the proof of Theorem 6.6.3 is provided at the end of the present subsection; it of course follows the methodology described in Section 6.3.

The bounds by Birgé [2005] only deal with events $A_1, \dots, A_N$ forming a partition of the underlying measurable space. As should be clear from their proof this assumption is crucial.

Theorem 6.6.3 (Birgé's lemma)**.** *Given an underlying measurable space* $(\Omega, \mathcal{F})$*, for all* $N \geqslant 2$*, for all probability distributions* $\mathbb{P}_1, \dots, \mathbb{P}_N$*, for all events* $A_1, \dots, A_N$ *forming a partition of* Ω*,*

$$\min_{1 \leqslant i \leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{c_N,\ \frac{\overline{K}}{\log(N)}\right\} \qquad \text{where} \qquad \overline{K} = \frac{1}{N-1} \sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1)$$

and where $(c_N)_{N \geqslant 2}$ *is a decreasing sequence, where each term* c_N *is defined as the unique* $c \in (0, 1)$ *such that*

$$\frac{-\big(c \log(c) + (1-c)\log(1-c)\big)}{c} + \log(1-c) = \log\left(\frac{N-1}{N}\right). \tag{6.36}$$

We have, for instance, $c_2 \approx 0.7587$ *and* $c_3 \approx 0.7127$*, while* $\lim c_N = 0.63987$*.*

The aim of this subsection is to compare this bound to the versions of Fano's inequality following from the kl lower bounds (6.11), (6.10), and (6.13), in this order. In the setting of the theorem above and by picking constant alternatives $\mathbb{Q}$, these lower bounds on kl respectively lead to

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(A_i) \leqslant \frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^{N}\mathrm{KL}(\mathbb{P}_i,\mathbb{Q}) + \log\left(2-\frac{1}{N}\right)}{\log(N)} \leqslant \frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^{N}\mathrm{KL}(\mathbb{P}_i,\mathbb{Q}) + \log(2)}{\log(N)}\,, \tag{6.37}$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(A_i) \leqslant \frac{1}{N} + \sqrt{\frac{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^{N}\mathrm{KL}(\mathbb{P}_i,\mathbb{Q})}{\max\{\log(N),\,2\}}}\,. \tag{6.38}$$

The main point of Birgé [2005] was that the most classical version of Fano's inequality, that is, the right-most side of (6.37), was quite unpractical for small values of N, and even useless when $N = 2$. In the latter case $N = 2$, the statistical doxa had it that one should rather resort to Pinsker's inequality, which is exactly (6.38) when $N = 2$. One of his main motivation was therefore to get an inequality that would be useful for all $N \geqslant 2$, so that one does not have to decide which of the classical Pinsker's inequality or the classical Fano's inequality should be applied. A drawback, however, of his bound is the $\overline{K}$ term, in which one cannot pick a convenient $\mathbb{Q}$ as in the bounds (6.37)–(6.38). Also, the result is about the minimum of the $\mathbb{P}_i(A_i)$, not about their average.

Now, we note that both the middle term in (6.37) and the bound (6.38) yield useful bounds, even for $N = 2$. The middle term in (6.37) was derived—with a different formulation—by Chen et al. [2016], see Proposition 6.6.1 above. Our contribution is to note that our inequality (6.38) provides an interpolation between Pinsker's and Fano's inequalities. More precisely, (6.38) implies both Pinsker's inequality and, lower bounding the maximum by $\log(N)$, a bound as useful as Theorem 6.6.3. Indeed, in practice, the additional additive $1/N$ term and the additional square root do not prevent from obtaining the desired lower bounds, as illustrated in Section 6.4.2.

We close this subsection with a proof of Theorem 6.6.3.

Proof (of Theorem 6.6.3): We denote by $h : p \in [0,1] \mapsto -\big(p\log(p)+(1-p)\log(1-p)\big)$ the binary entropy function. The existence of c_N follows from the fact that $c \in (0,1) \mapsto h(c)/c + \log(1-c)$ is continuous and decreasing, as the sum of two such functions; its respective limits are $+\infty$ and $-\infty$ at 0 and 1.

Reduction (6.2) with $\mathbb{Q}_i = \mathbb{P}_1$ for all $i \geqslant 2$ indicates that $\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) \leqslant \overline{K}$ where

$$\widetilde{p} \overset{\text{def}}{=} \frac{1}{N-1} \sum_{i=2}^{N} \mathbb{P}_i(A_i)\,, \qquad \widetilde{q} \overset{\text{def}}{=} \frac{1}{N-1} \sum_{i=2}^{N} \mathbb{P}_1(A_i) = \frac{1 - \mathbb{P}_1(A_1)}{N-1}\,,$$
$$\overline{K} = \frac{1}{N-1} \sum_{i=2}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_1)\,;$$

note that we used the assumption of a partition to get the alternative definition of the $\widetilde{q}$ quantity. We use the following lower bound on kl, which follows from calculations similar to the ones performed in (6.4), using that $c_N \geqslant 1/2$ and that the binary entropy $h : p \mapsto -\big(p \log(p) + (1-p)\log(1-p)\big)$ is decreasing on $[1/2,\, 1]$: for $p \geqslant c_N$,

$$\mathrm{kl}(p, q) \geqslant p \log\left(\frac{1}{q}\right) - h(c_N) \geqslant p \log\left(\frac{1}{q}\right) - p\,\frac{h(c_N)}{c_N}\,,$$

where $\log(1/q) - h(c_N)/c_N > 0$ for $q < \exp\big(-h(c_N)/c_N\big)$. Hence,

$$\forall p \in [0,1], \quad \forall\, q \in \Big(0,\, \exp\big(-h(c_N)/c_N\big)\Big), \qquad p \leqslant \max\left\{c_N,\ \frac{\mathrm{kl}(p,q)}{\log(1/q) - h(c_N)/c_N}\right\}. \tag{6.39}$$

Now, we set $a = \min\limits_{1 \leqslant i \leqslant N} \mathbb{P}_i(A_i)$ and may assume $a \geqslant c_N$ (otherwise, the stated bound is obtained).

We have, by the very definition of a as a minimum and by the definition (6.36) of c_N,

$$a \leqslant \widetilde{p} \qquad \text{and} \qquad \widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant \frac{1-c_N}{N-1} = \frac{1}{N} \exp\left(-\frac{h(c_N)}{c_N}\right), \tag{6.40}$$

while $\widetilde{q} > 0$ unless $\mathbb{P}_1(A_1) = a = 0$, in which case there is nothing to prove. We may therefore combine $\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big) \leqslant \overline{K}$ with (6.39) to get

$$a \leqslant \widetilde{p} \leqslant \max\left\{c_N,\ \frac{\mathrm{kl}\big(\widetilde{p}, \widetilde{q}\big)}{\log\big(1/\widetilde{q}\big) - h(c_N)/c_N}\right\} \leqslant \max\left\{c_N,\ \frac{\overline{K}}{\log(N)}\right\},$$

where, for the last inequality, we used the upper bound on $\widetilde{q}$ in (6.40). □

6.7 Proofs of the stated lower bounds on kl (and of an improved Bretagnolle-Huber inequality)

We prove in this section the convexity inequalities (6.11) and (6.12) as well as the refined Pinsker's inequality and its consequence (6.13). Using the same techniques and methodology as for establishing these bounds, we also improve in passing the Bretagnolle-Huber inequality.

6.7.1 Proofs of the convexity inequalities (6.11) and (6.12)

Proof. Inequality (6.12) follows from (6.11) via a function study of $q \in (0,1) \mapsto \log(2-q)\big/\log(1/q)$, which is dominated by $0.21 + 0.79\,\mathrm{q}$.

Now, the shortest proof of (6.11) notes that the duality formula for the Kullback-Leibler divergence between Bernoulli distributions—already used in (6.27)—ensures that, for all $p \in [0,1]$ and $q \in (0,1]$,

$$\mathrm{kl}(p,q) = \sup_{\lambda \in \mathbb{R}} \left\{ \lambda p - \log\Big(q\big(\mathrm{e}^{\lambda} - 1\big) + 1 \Big) \right\} \geqslant p \log\left(\frac{1}{q}\right) - \log(2-q)$$

for the choice $\lambda = \log(1/q)$. □

An alternative, longer but more elementary proof uses a direct convexity argument, as in Guntuboyina [2011, Example II.4], which already included the inequality of interest in the special case when $q = 1/N$; see also Chen et al. [2016]. We deal separately with $p = 0$ and $p = 1$, and thus restrict our attention to $p \in (0,1)$ in the sequel. For $q \in (0,1)$, as $p \mapsto \mathrm{kl}(p,q)$ is convex and differentiable on $(0,1)$, we have

$$\forall\, (p, p_0) \in (0,1)^2, \qquad \mathrm{kl}(p,q) - \mathrm{kl}(p_0,q) \geqslant \underbrace{\log\left(\frac{p_0(1-q)}{(1-p_0)q}\right)}_{\frac{\partial}{\partial p}\mathrm{kl}(p_0,q)} (p - p_0)\,. \tag{6.41}$$

The choice $p_0 = 1/(2-q)$ is such that

$$\frac{p_0}{1-p_0} = \frac{1}{1-q}\,, \qquad \text{thus} \qquad \log\left(\frac{p_0(1-q)}{(1-p_0)q}\right) = \log\left(\frac{1}{q}\right),$$

and

$$\mathrm{kl}(p_0,q) = \frac{1}{2-q}\log\left(\frac{1/(2-q)}{q}\right) + \frac{1-q}{2-q}\log\left(\frac{(1-q)/(2-q)}{1-q}\right) = \frac{1}{2-q}\log\left(\frac{1}{q}\right) + \log\left(\frac{1}{2-q}\right).$$

Inequality (6.41) becomes

$$\forall\, p \in (0,1), \qquad \mathrm{kl}(p,q) - \frac{1}{2-q}\log\left(\frac{1}{q}\right) + \log(2-q) \geqslant \left(p - \frac{1}{2-q}\right)\log\left(\frac{1}{q}\right),$$

which proves as well the bound (6.11).

6.7.2 Proofs of the refined Pinsker's inequality and of its consequence (6.13)

The next theorem is a stronger version of Pinsker's inequality for Bernoulli distributions, that was proved[2] by Ordentlich and Weinberger [2005]. Indeed, note that the function φ

[2] We also refer the reader to Kearns and Saul [1998, Lemma 1] and Berend and Kontorovich [2013, Theorem 3.2] for dual inequalities upper bounding the moment-generating function of the Bernoulli distributions.

defined below satisfies $\min \varphi = 2$, so that the next theorem always yields an improvement over the most classical version of Pinsker's inequality: $\mathrm{kl}(p,q) \geqslant 2(p-q)^2$.

We provide below an alternative elementary proof for Bernoulli distributions of this refined Pinsker's inequality. The extension to the case of general distributions, via the contraction-of-entropy property, is stated at the end of this section.

Theorem 6.7.1 (A refined Pinsker's inequality by Ordentlich and Weinberger [2005])**.** *For all $p, q \in [0,1]$,*

$$\mathrm{kl}(p,q) \geqslant \frac{\log\bigl((1-q)/q\bigr)}{1-2q}\,(p-q)^2 \stackrel{\mathrm{def}}{=} \varphi(q)\,(p-q)^2\,,$$

where the multiplicative factor $\varphi(q) = (1-2q)^{-1}\log\bigl((1-q)/q\bigr)$ is defined for all $q \in [0,1]$ by extending it by continuity as $\varphi(1/2) = 2$ and $\varphi(0) = \varphi(1) = +\infty$.

The proof shows that $\varphi(q)$ is the optimal multiplicative factor in front of $(p-q)^2$ when the bounds needs to hold for all $p \in [0,1]$; the proof also provides a natural explanation for the value of φ.

Proof. The stated inequality is satisfied for $q \in \{0,1\}$ as $\mathrm{kl}(p,q) = +\infty$ in these cases unless $p = q$. The special case $q = 1/2$ is addressed at the end of the proof. We thus fix $q \in (0,1) \setminus \{1/2\}$ and set $f(p) = \mathrm{kl}(p,q)/(p-q)^2$ for $p \neq q$, with a continuity extension at $p = q$. We exactly show that f attains its minimum at $p = 1-q$, from which the result (and its optimality) follow by noting that

$$f(1-q) = \frac{\mathrm{kl}(1-q,q)}{(1-2q)^2} = \frac{\log\bigl((1-q)/q\bigr)}{1-2q} = \varphi(q)\,.$$

Given the form of f, it is natural to perform a second-order Taylor expansion of $\mathrm{kl}(p,q)$ around q. We have

$$\frac{\partial}{\partial p}\mathrm{kl}(p,q) = \log\left(\frac{p(1-q)}{(1-p)q}\right) \qquad \text{and} \qquad \frac{\partial^2}{\partial^2 p}\mathrm{kl}(p,q) = \frac{1}{p(1-p)} \stackrel{\mathrm{def}}{=} \psi(p)\,, \tag{6.42}$$

so that Taylor's formula with integral remainder reveals that for $p \neq q$,

$$f(p) = \frac{\mathrm{kl}(p,q)}{(p-q)^2} = \frac{1}{(p-q)^2}\int_q^p \frac{\psi(t)}{1!}(p-t)^1 \mathrm{d}t = \int_0^1 \psi\bigl(q+u(p-q)\bigr)(1-u)\mathrm{d}u\,.$$

This rewriting of f shows that f is strictly convex (as ψ is so). Its global minimum is achieved at the unique point where its derivative vanishes. But by differentiating under the integral sign, we have, at $p = 1-q$,

$$f'(1-q) = \int_0^1 \psi'\bigl(q+u(1-2q)\bigr)\,u(1-u)\mathrm{d}u = 0\,;$$

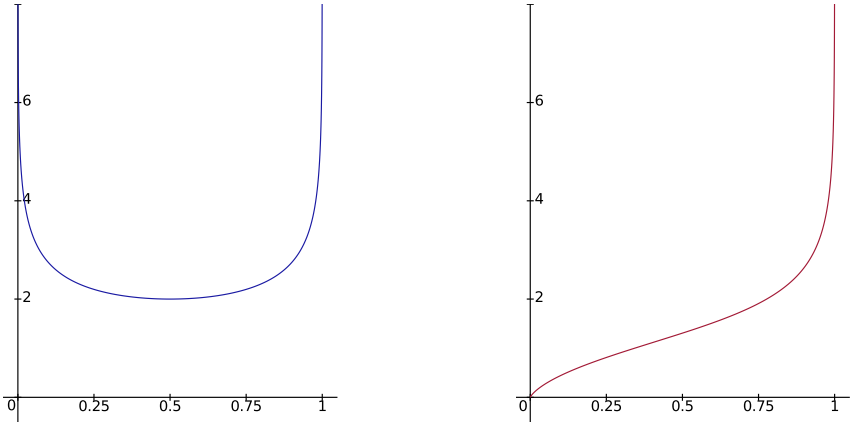

Figure 6.1: Plots of φ [left] and $x \in (0,1) \mapsto \varphi(x) - \log(1/x)$ [right].

the equality to 0 follows from the fact that the function $u \mapsto \psi'\big(q + u(1-2q)\big)u(1-u)$ is antisymmetric around $u = 1/2$ (essentially because ψ' is antisymmetric itself around $1/2$). As a consequence, the convex function f attains its global minimum at $1-q$, which concludes the proof for the case where $q \in (0,1) \setminus \{1/2\}$.

It only remains to deal with $q = 1/2$: we use the continuity of $\mathrm{kl}(p, \cdot\,)$ and φ to extend the obtained inequality from $q \in [0,1] \setminus \{1/2\}$ to $q = 1/2$. □

We now prove the second inequality of (6.13). A picture is helpful, see Figure 6.1.

Corollary 6.7.2. *For all $q \in (0,1]$, we have $\varphi(q) \geqslant 2$ and $\varphi(q) \geqslant \log(1/q)$. Thus, for all $p \in [0,1]$ and $q \in (0,1)$,*

$$p \leqslant q + \sqrt{\frac{\mathrm{kl}(p,q)}{\max\big\{\log(1/q),\, 2\big\}}}\,.$$

Slightly sharper bounds are possible, like $\varphi(q) \geqslant (1+q)(1+q^2)\log(1/q)$ or $\varphi(q) \geqslant \log(1/q) + 2.5\,q$, but we were unable to exploit these refinements in our applications.

General refined Pinsker's inequality. The following result, which improves on Pinsker's inequality, is due to Ordentlich and Weinberger [2005]. Our approach through Bernoulli distributions enables to derive it in an elementary (and enlightening) way: by combining Theorem 6.7.1 and the data-processing inequality (Lemma 6.2.2).

Theorem 6.7.3. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$. Then*

$$\sup_{A \in \mathcal{F}} \big|\mathbb{P}(A) - \mathbb{Q}(A)\big| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{\inf\limits_{A \in \mathcal{F}} \varphi\big(\mathbb{Q}(A)\big)}}\,,$$

where $\varphi \geqslant 2$ is defined in the statement of Theorem 6.7.1.

6.7.3 An improved Bretagnolle-Huber inequality

The Bretagnolle-Huber inequality was introduced by Bretagnolle and Huber [1978, 1979]. The multiplicative factor $\mathrm{e}^{-1/\mathrm{e}} \geqslant 0.69$ in our statement (6.43) below is a slight improvement over the original $1/2$ factor. For all $p, q \in [0,1]$,

$$1 - |p-q| \geqslant \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\mathrm{kl}(p,q)}\,, \qquad \text{thus} \qquad q \geqslant p - 1 + \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\mathrm{kl}(p,q)}\,. \tag{6.43}$$

It is worth to note that Bretagnolle and Huber [1978] also proved the inequality

$$|p-q| \leqslant \sqrt{1 - \exp\bigl(-\mathrm{kl}(p,q)\bigr)}\,,$$

which improves as well upon the Bretagnolle-Huber inequality with the $1/2$ factor, but which is neither better nor worse than (6.43).

Now, via the data-processing inequality (Lemma 6.2.2), we get from (6.43)

$$1 - \sup_{A \in \mathcal{F}} \bigl|\mathbb{P}(A) - \mathbb{Q}(A)\bigr| \geqslant \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\mathrm{KL}(\mathbb{P},\mathbb{Q})}\,.$$

The left-hand side can be rewritten as $\inf_{A \in \mathcal{F}} \bigl\{\mathbb{P}(A) + \mathbb{Q}(A^c)\bigr\}$, where A^c denotes the complement of A. Therefore, the above inequality is a lower bound on the test affinity between $\mathbb{P}$ and $\mathbb{Q}$. For the sake of comparison to (6.25), we can restate the general version of the Bretagnolle-Huber inequality as: for all $A \in \mathcal{F}$,

$$\mathbb{Q}(A) \geqslant \mathbb{P}(A) - 1 + \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\mathrm{KL}(\mathbb{P},\mathbb{Q})}\,. \tag{6.44}$$

We now provide a proof of (6.43); note that our improvement was made possible because we reduced the proof to very elementary arguments in the case of Bernoulli distributions.

Proof. The case where $p \in \{0,1\}$ or $q \in \{0,1\}$ can be handled separately; we consider $(p,q) \in (0,1)^2$ in the sequel. The derivative of the function $x \in (0,1) \mapsto x \log\bigl(x/(1-q)\bigr)$ equals $1 + \log(x) - \log(1-q)$, so that the function achieves its minimum at $x = (1-q)/\mathrm{e}$, with value $-(1-q)/\mathrm{e} \geqslant -1/\mathrm{e}$. Therefore,

$$-\mathrm{kl}(p,q) = -p \log\left(\frac{p}{q}\right) - (1-p) \log\left(\frac{1-p}{1-q}\right) \leqslant -p \log\left(\frac{p}{q}\right) + \frac{1}{\mathrm{e}} = p\left(\log\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\frac{1}{\mathrm{e}}\,.$$

Therefore, using the convexity of the exponential,

$$\mathrm{e}^{-\mathrm{kl}(p,q)} \leqslant p\, \exp\left(\log\left(\frac{q}{p}\right) + \frac{1}{\mathrm{e}}\right) + (1-p)\, \mathrm{e}^{1/\mathrm{e}} = \bigl(q + (1-p)\bigr)\, \mathrm{e}^{1/\mathrm{e}}\,,$$

which shows that

$$1 - (p-q) \geqslant \mathrm{e}^{-1/\mathrm{e}}\, \mathrm{e}^{-\mathrm{kl}(p,q)}\,.$$

By replacing q by $1-q$ and p by $1-p$, we also get

$$1-(q-p) = 1-\big((1-p)-(1-q)\big) \geqslant \mathrm{e}^{-1/\mathrm{e}}\,\mathrm{e}^{-\mathrm{kl}(1-p,1-q)} = \mathrm{e}^{-1/\mathrm{e}}\,\mathrm{e}^{-\mathrm{kl}(p,q)}\,.$$

This concludes the proof, as $1-|p-q|$ is equal to the smallest value between $1-(p-q)$ and $1-(q-p)$. □

6.8 Elements of Proof

6.8.1 Two toy applications of the continuous Fano's inequality

We present here two toy applications of our continuous Fano's inequality, that the unfamiliar reader may study in complement to the new applications addressed in Sections 6.4 and 6.5. The two topics covered below are:

- parametric density estimation in the multivariate Gaussian model, where we use the reduction (6.7);
- nonparametric regression with fixed design, which also relies on the reduction (6.7).

Parametric minimax lower bound in the mutivariate Gaussian model

The next result is a well-known minimax lower bound on the mean-estimation problem in the standard multivariate Gaussian model. Many proof techniques were used to derive this toy lower bound. The proof we provide below illustrates how to use Fano's inequality without relying on any discretization argument, thanks to a continuous version of Fano's inequality. Proofs of the same spirit were proposed by Duchi and Wainwright [2013] and Chen et al. [2016], though some discretizations were still used at some point in both references.

See the end of this subsection for extended comments and references, in particular with respect to other well-known proof techniques not resorting to discretization arguments like the use of Assouad's lemma.

Proposition 6.8.1 (Parametric lower bound with a continuous Fano's inequality)**.** *Let $d \geqslant 1$ be the ambient dimension and $\mathbb{R}^d$ be the parameter space. Assume that we observe an n–sample $X_1, \dots, X_n$ distributed according to $\mathcal{N}(\theta, \sigma^2 I_d)$ for some unknown $\theta \in \mathbb{R}^d$, where $\sigma > 0$ and where I_d is the $d \times d$ identity matrix. Then, denoting by $\mathbb{E}_\theta$ the expectation when the unknown parameter is θ, we have the lower bound*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta \Big[\big\| \widehat{\theta} - \theta \big\|_2^2 \Big] \geqslant c_d \, \frac{\sigma^2 d}{n} \, ,$$

where the infimum is over all $\mathbb{R}^d$–valued estimators $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n)$, and where $(c_d)_{d \geqslant 1}$ is an increasing sequence such that $c_1 \geqslant 0.01$, $c_2 \geqslant 0.025$, and $\lim_{d \to +\infty} c_d \geqslant 0.05$.

The proof of this result uses similar but simpler arguments than the one of Proposition 6.4.1.

Proof. We may assume with no loss of generality that the underlying probability space is $\Omega = \big(\mathbb{R}^d\big)^n$, that each X_i is the i–th projection map $(x_1, \dots, x_n) \in \Omega \mapsto x_i \in \mathbb{R}^d$, and that the collection of probability distributions over Ω is formed by the $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2 I_d)^{\otimes n}$. We still denote by $\mathbb{E}_\theta$ the expectation under $\mathbb{P}_\theta$. Fix any estimator $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n)$ and

let $\varepsilon > 0$ be determined by the analysis. By Markov's inequality for the first inequality and considering any probability distribution ν over $\mathbb{R}^d$ for the second inequality,

$$\sup_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\Big[\big\|\widehat{\theta}-\theta\big\|_2^2\Big] \geqslant \sup_{\theta\in\mathbb{R}^d} \varepsilon^2\,\mathbb{P}_\theta\Big(\big\|\widehat{\theta}-\theta\big\|_2^2 > \varepsilon^2\Big) = \varepsilon^2\left(1-\inf_{\theta\in\mathbb{R}^d}\mathbb{P}_\theta\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\right)$$
$$\geqslant \varepsilon^2\left(1-\int_{\mathbb{R}^d}\mathbb{P}_\theta\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\mathrm{d}\nu(\theta)\right). \tag{6.45}$$

We take ν as the uniform distribution on the Euclidean ball $B(0,\rho\varepsilon) = \big\{u\in\mathbb{R}^d : \|u\|_2\leqslant\rho\varepsilon\big\}$ for some $\rho > 1$ to be determined by the analysis (as is also the case for ε). Fano's inequality in the form given by the combination of (6.7) and (6.13), with the fixed alternative distribution $\mathbb{P}_0$ (where 0 denotes the null vector of $\mathbb{R}^d$) and the sets $A_\theta = \big\{\|\widehat{\theta}-\theta\|_2\leqslant\varepsilon\big\}$, indicates that

$$\int_{B(0,\rho\varepsilon)}\mathbb{P}_\theta\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\mathrm{d}\nu(\theta)$$
$$\leqslant \int_{B(0,\rho\varepsilon)}\mathbb{P}_0\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\mathrm{d}\nu(\theta) + \sqrt{\frac{\displaystyle\int_{B(0,\rho\varepsilon)}\mathrm{KL}\big(\mathbb{P}_\theta,\mathbb{P}_0\big)\mathrm{d}\nu(\theta)}{\displaystyle-\log\left(\int_{B(0,\rho\varepsilon)}\mathbb{P}_0\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\mathrm{d}\nu(\theta)\right)}}$$
$$\leqslant \left(\frac{1}{\rho}\right)^d + \sqrt{\frac{n\rho^2\varepsilon^2/(2\sigma^2)}{d\log\rho}}\,, \tag{6.46}$$

where the second inequality follows from the inequalities (6.47) and (6.48) below. First note that, by independence, $\mathrm{KL}\big(\mathbb{P}_\theta,\mathbb{P}_0\big) = n\mathrm{KL}\big(\mathcal{N}(\theta,\sigma^2),\mathcal{N}(0,\sigma^2)\big) = n\,\|\theta\|_2^2/(2\sigma^2)$, so that

$$\int_{B(0,\rho\varepsilon)}\mathrm{KL}\big(\mathbb{P}_\theta,\mathbb{P}_0\big)\mathrm{d}\nu(\theta) = \frac{n}{2\sigma^2}\int_{B(0,\rho\varepsilon)}\|\theta\|_2^2\,\mathrm{d}\nu(\theta) \leqslant \frac{n\rho^2\varepsilon^2}{2\sigma^2}\,. \tag{6.47}$$

Second, by the Fubini-Tonelli theorem,

$$q \overset{\text{def}}{=} \int_{B(0,\rho\varepsilon)}\mathbb{P}_0\Big(\big\|\widehat{\theta}-\theta\big\|_2\leqslant\varepsilon\Big)\mathrm{d}\nu(\theta) = \mathbb{E}_0\left[\int_{B(0,\rho\varepsilon)}\mathbb{1}_{\left\{\|\widehat{\theta}-\theta\|_2\leqslant\varepsilon\right\}}\mathrm{d}\nu(\theta)\right]$$
$$= \mathbb{E}_0\Big[\nu\Big(B\big(\widehat{\theta},\varepsilon\big)\cap B\big(0,\rho\varepsilon\big)\Big)\Big] \leqslant \left(\frac{1}{\rho}\right)^d, \tag{6.48}$$

where to get the last inequality we used the fact that, almost surely, $\nu\big(B\big(\widehat{\theta},\varepsilon\big)\cap B\big(0,\rho\varepsilon\big)\big)$ is the ratio of the volume of a (possibly truncated) Euclidean ball of radius ε with the volume of the support of ν, namely, the larger Euclidean ball $B(0,\rho\varepsilon)$, in dimension d.

We conclude the proof by combining (6.45) and (6.46) and recalling that $\rho > 1$ and $\varepsilon > 0$ were two parameters to get

$$\sup_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\Big[\big\|\widehat{\theta} - \theta\big\|_2^2\Big] \geqslant \sup_{\rho>1}\sup_{\varepsilon>0} \varepsilon^2\left(1 - \left(\frac{1}{\rho}\right)^d - \sqrt{\frac{n\rho^2\varepsilon^2}{2\sigma^2 d\log\rho}}\right)$$
$$= \frac{d\sigma^2}{n}\underbrace{\sup_{\rho>1}\frac{8\Big(1-(1/\rho)^d\Big)^3\log(\rho)}{27\rho^2}}_{=c_d}$$

for the optimal choice of $\varepsilon = (2/3)\big(1-(1/\rho)^d\big)\sqrt{2d\sigma^2\log(\rho)/(n\rho^2)}$. We see $c_1 \geqslant 0.01$ and $c_2 \geqslant 0.025$ via the respective choices $\rho = 4$ and $\rho = 2.5$, while the fact that the limit is larger than 0.05 follows, e.g., from the choice $\rho = 2$.

Note that, when using (6.13) above, we implicitly assumed that the quantity q defined in (6.48) lies in $(0,1)$. The fact that $q < 1$ follows directly from the upper bound $(1/\rho)^d$ given $\rho > 1$. As for the condition $q > 0$, note that given the support of ν, we could rewrite (6.45) as

$$\sup_{\theta\in\mathbb{R}^d} \mathbb{E}_\theta\Big[\big\|\widehat{\theta} - \theta\big\|_2^2\Big] \geqslant \sup_{\theta\in B(0,\rho\varepsilon)} \mathbb{E}_\theta\Big[\big\|\widehat{\theta} - \theta\big\|_2^2\Big] \geqslant \sup_{\theta\in B(0,\rho\varepsilon)} \mathbb{E}_\theta\Big[\big\|\Pi(\widehat{\theta}) - \theta\big\|_2^2\Big]$$
$$\geqslant \varepsilon^2\left(1 - \int_{\mathbb{R}^d} \mathbb{P}_\theta\Big(\big\|\Pi(\widehat{\theta}) - \theta\big\|_2 \leqslant \varepsilon\Big)\mathrm{d}\nu(\theta)\right),$$

where $\Pi(\widehat{\theta})$ is the projection of $\widehat{\theta}$ onto the closed convex set $B\big(0,\rho\varepsilon\big)$. Thus, we can assume without loss of generality that $\widehat{\theta} \in B(0,\rho\varepsilon)$ almost surely. In this case, the normalized volume $\nu\big(B\big(\widehat{\theta},\varepsilon\big)\cap B\big(0,\rho\varepsilon\big)\big)$ is almost surely positive, so that $q = \mathbb{E}_0\big[\nu\big(B\big(\widehat{\theta},\varepsilon\big)\cap B\big(0,\rho\varepsilon\big)\big)\big] > 0$. □

Comparison with other, historical proofs. Various types of proofs were proposed in the literature to derive a lower bound of order $d\sigma^2/n$ as above.

The proof technique that consists in lower bounding the minimax risk by the Bayes risk works surprisingly well in this simple estimation problem. It is indeed folklore knowledge that taking a Gaussian prior with covariance matrix $s^2 I_d$ and letting $s \to +\infty$ yields, after simple calculations, a lower bound of $d\sigma^2/n$ (see, e.g., Massart, 2007, page 106). Interestingly the multiplicative constant of 1 is even optimal because it matches the upper bound of $d\sigma^2/n$ satisfied by the empirical mean. However, this proof technique does not carry over easily to more complex settings such as, e.g., the same Gaussian model but with a bounded parameter space Θ, as is the case in the nonparametric regression problem of Section 6.8.1 below. This is the reason why, even for this toy estimation problem, it is useful to provide alternative proofs that may be suboptimal in terms of the multiplicative constant, but that can be easily adapted to more intricate settings.

Another simple proof technique consists in using Assouad's lemma, which is very useful when the loss function can be decomposed as a sum over the d coordinates, as is the case here. Assouad's lemma reduces the estimation problem to d parallel two-hypotheses testing problems. (See, e.g., Yu, 1997, Example 2 for an application of Assouad's lemma.)

All alternative proofs that we know of are based on Fano's inequality and often involve a discretization argument. Historical proofs reduce the estimation problem to a multiple-hypotheses testing problem (with exponentially many hypotheses) by showing that every estimator must fail to identify at least one θ in a subset of the d–dimensional rescaled hypercube $\{0, c\sigma/\sqrt{n}\}^d$ (this subset is obtained via a combinatorial tool known as Varshamov-Gilbert's lemma). See, e.g., Yu [1997, Example 2] or Massart [2007, Proposition 4.8] for such applications of Fano's inequality. More recently Duchi and Wainwright [2013] and Chen et al. [2016] provided a continuous version of Fano's inequality (of which Lemma 6.3.3 above is a generalization to some extent) to avoid the discretization step mentioned earlier. Instead they directly addressed a multiple testing problem with continuously many hypotheses. This provides a nice interpretation of the factor d in the lower bound $d\sigma^2/n$ as the log ratio of the volumes of two Euclidean balls in $\mathbb{R}^d$, as in (6.48) above. Note however that both articles use a discretization argument at some point: Duchi and Wainwright [2013] prove their continous Fano inequality (Proposition 2 therein) via an unnecessarily involved grid-based approximation argument. Chen et al. [2016] later proved a continuous Fano's inequality (cf. Corollary 3.5 and Theorem 4.1 therein) without any discretization argument, but the way they use it in Example 5.3 for the Gaussian model relies on an unnecessary calculation of covering numbers. On the contrary, the proof we provided above uses no discretization whatsoever.

We finally mention the lower bound that Xu and Raginsky [2016] derived for the Bayes risk with a uniform prior on a Euclidean ball (as in the proof above). The proof of their Corollary 3, which uses a generalized Fano's inequality, also bypasses any discretization step. It is however only asymptotic in n, and it requires longer calculations than above since log ratios of densities have to be manipulated explicitly.

A minimax lower bound for nonparametric regression

In this subsection we revisit a well-known lower bound within the nonparametric regression model with fixed design, which unfolds as follows. We observe an n–sample $(x_1, Y_1), \dots, (x_n, Y_n)$, where $x_i = (i-1)/n \in [0,1]$ and

$$Y_i = f(x_i) + \varepsilon_i \;, \quad \text{where} \quad 1 \leqslant i \leqslant n \,,$$

for some unknown function $f \in \mathcal{F}_L \stackrel{\text{def}}{=} \big\{g : \; [0,1] \to \mathbb{R}, \; g \text{ is } L\text{–Lipschitz}\big\}$ and ε_i that are independent and identically distributed according to $\mathcal{N}(0, \sigma^2)$. The goal is to estimate f; the parameters are σ^2, L and f. For the sake of notation, we only focus on f and denote by $\mathbb{P}_f$ and $\mathbb{E}_f$ the probability and expectation underlying the random vector

$(Y_1, \ldots, Y_n)$. Actually, as in the previous section and with no loss of generality, we may identify the law of $(Y_1, \ldots, Y_n)$ and the underlying probability.

We assess the accuracy of any estimator $\widehat{f} \in \mathbb{L}^2\big([0,1]\big)$ via its expected quadratic risk,

$$\mathbb{E}_f\Big[\big\|\widehat{f} - f\big\|_2^2\Big],$$

where the squared Euclidean norm of any $g \in \mathbb{L}^2\big([0,1]\big)$ is defined as $\|g\|_2^2 = \displaystyle\int_0^1 g(x)^2 \mathrm{d}x$.

The next lower bound is well known; see Ibragimov and Has'minskii [1982, 1984], Tsybakov [2009, Theorem 2.8] or Duchi [2014, Theorem 4.4] for proofs based on either Fano's inequality (with a discretization argument) or Assouad's lemma. We illustrate below how to use the continuous Fano's inequality.

Proposition 6.8.2 (Nonparametric lower bound with no discretization)**.** *Fix $\sigma^2 > 0$ and $L > 0$, two quantities possibly known to the statistician. In the nonparametric regression model described above, we have, for all $n \geqslant \max\big\{L/\big(\sigma\sqrt{\log(2)}\big),\, 64\log(2)\,\sigma^2/L^2\big\}$,*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_L} \mathbb{E}_f\Big[\big\|\widehat{f} - f\big\|_2^2\Big] \geqslant C\left(\frac{\sigma^2 L}{n}\right)^{2/3},$$

where the infimum is over all estimators $\widehat{f} = \widehat{f}(Y_1, \ldots, Y_n) \in \mathbb{L}^2\big([0,1]\big)$, and where C is a universal positive constant; e.g., $C = 0.001$ works.

The lower bound of Proposition 6.8.2 is tight in the sense that there exist estimators $\widehat{f}$ such that

$$\sup_{f \in \mathcal{F}_L} \mathbb{E}_f\Big[\big\|\widehat{f} - f\big\|_2^2\Big] \leqslant \widetilde{C}\left(\frac{\sigma^2 L}{n}\right)^{2/3}$$

for some universal constant $\widetilde{C}$; see, e.g., Tsybakov [2009, Theorem 1.7] or Duchi [2014, Corollary 4.3].

Proof. We start as in the proof of Duchi [2014, Theorem 4.4]. Let φ be the function defined on $\mathbb{R}$ by $\varphi(x) = \big(1/2 - |x - 1/2|\big)_+$, where $y_+ = \max\{0, y\}$. Note that φ is a 1–Lipschitz function with support $(0,1)$; in addition,

$$\|\varphi\|_2^2 = 2\int_0^{1/2} x^2 \mathrm{d}x = \frac{1}{12}.$$

Thus, for any integer $d \geqslant 2$ (to be determined by the analysis), the functions $f_j : [0,1] \to \mathbb{R}$ defined for all $j \in \{1, \ldots, d\}$ by

$$f_j(x) = \frac{L}{d}\,\varphi\left(d\left(x - \frac{j-1}{d}\right)\right)$$

belong to $\mathcal{F}_L$ and form an orthogonal system in $\mathbb{L}^2\big([0,1]\big)$, since the f_j have pairwise disjoint supports. We define $f_\theta = \sum_{j=1}^d \theta_j f_j$ for all $\theta \in \mathbb{R}^d$. Note that, again because of the disjoint supports of the f_j, the mapping $\theta \mapsto f_\theta$ is an injection from Θ into $\mathcal{F}_L$, where

$$\Theta \stackrel{\text{def}}{=} \big\{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leqslant 1\big\}\,.$$

Moreover, for all $\theta, \theta' \in \Theta$, we have the norm relationship

$$\|f_\theta - f_{\theta'}\|_2^2 = \sum_{j=1}^d (\theta_j - \theta'_j)^2\, \|f_j\|_2^2 = \frac{L^2}{d^3}\, \|\varphi\|_2^2\, \|\theta - \theta'\|_2^2 = \frac{L^2}{12\, d^3} \|\theta - \theta'\|_2^2\,. \tag{6.49}$$

Next we reduce the nonparametric problem to a parametric one and then proceed as in the proof of Proposition 6.8.1, avoiding any discretization argument. To that end, we write abusively $\mathbb{P}_\theta = \mathbb{P}_{f_\theta}$ and $\mathbb{E}_\theta = \mathbb{E}_{f_\theta}$.

We set $\widehat{\theta} = \arg\min_{\theta\in\Theta} \|\widehat{f} - f_\theta\|_2^2$ and get

$$\sup_{f\in\mathcal{F}_L} \mathbb{E}_f\Big[\big\|\widehat{f} - f\big\|_2^2\Big] \geqslant \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\big\|\widehat{f} - f_\theta\big\|_2^2\Big] \geqslant \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\big\|f_{\widehat{\theta}} - f_\theta\big\|_2^2\Big]\,, \tag{6.50}$$

where we first used that the set $\mathcal{F}_\Theta = \big\{f_\theta : \theta \in \Theta\big\}$ is a subset of $\mathcal{F}_L$, and second, that $f_{\widehat{\theta}}$ is the projection of $\widehat{f}$ onto the closed convex subset $\mathcal{F}_\Theta$ of $\mathbb{L}^2\big([0,1]\big)$. Now, for all $\rho > 0$ (to be determined by the analysis), inequality (6.49) and Markov's inequality yield

$$\begin{aligned}
&\sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\big\|f_{\widehat{\theta}} - f_\theta\big\|_2^2\Big] \\
&= \frac{L^2}{12\, d^3} \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\big\|\widehat{\theta} - \theta\big\|_2^2\Big] \geqslant \frac{L^2}{12\, d^3} \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\rho d\, \mathbb{1}_{\{\|\widehat{\theta}-\theta\|_2^2 \geqslant \rho d\}}\Big] \\
&\geqslant \frac{\rho L^2}{12\, d^2} \Big(1 - \inf_{\theta\in\Theta} \mathbb{P}_\theta\Big(\big\|\widehat{\theta} - \theta\big\|_2^2 \leqslant \rho d\Big)\Big) \geqslant \frac{\rho L^2}{12\, d^2} \Big(1 - \int_\Theta \mathbb{P}_\theta\Big(\big\|\widehat{\theta} - \theta\big\|_2^2 \leqslant \rho d\Big) \mathrm{d}\nu(\theta)\Big)\,,
\end{aligned} \tag{6.51}$$

where the last inequality holds true for any prior ν on Θ.

Now we choose the uniform (Lebesgue) prior $\nu \stackrel{\text{def}}{=} \mathcal{U}(\Theta)$ and apply Fano's inequality in the form of Lemma 6.3.3, with the fixed alternative distribution $\mathbb{P}_0$ (where 0 denotes the null vector of $\mathbb{R}^d$):

$$\int_\Theta \mathbb{P}_\theta\Big(\big\|\widehat{\theta} - \theta\big\|_2^2 \leqslant \rho d\Big) \mathrm{d}\nu(\theta) \leqslant \frac{\log(2) + \displaystyle\int_\Theta \mathrm{KL}\big(\mathbb{P}_\theta, \mathbb{P}_0\big) \mathrm{d}\nu(\theta)}{-\log\Big(\displaystyle\int_\Theta \mathbb{P}_0\Big(\big\|\widehat{\theta} - \theta\big\|_2^2 \leqslant \rho d\Big) \mathrm{d}\nu(\theta)\Big)} \leqslant \frac{1}{3} + \frac{(n+d)L^2}{8\, d^3 \sigma^2 \log(2)}\,, \tag{6.52}$$

where the second inequality follows from the inequalities (6.53) and (6.55) below, with the choice of $\rho = 1/(2\pi\mathrm{e})$. Note that these calculations also show that the integral in the denominator in (6.52) lies in $(0,1)$, as required for Lemma 6.3.3.

First note that $\mathbb{P}_\theta = \bigotimes_{i=1}^n \mathcal{N}\big(f_\theta(x_i), \sigma^2\big)$ and $\mathbb{P}_0 = \mathcal{N}(0,\sigma^2)^{\otimes n}$, so that, by independence,

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0) = \sum_{i=1}^n \mathrm{KL}\Big(\mathcal{N}\big(f_\theta(x_i), \sigma^2\big), \mathcal{N}(0,\sigma^2)\Big) = \frac{1}{2\sigma^2}\sum_{i=1}^n f_\theta(x_i)^2\,,$$

which, since the f_j have pairwise disjoint supports, is also equal to

$$\sum_{i=1}^n f_\theta(x_i)^2 = \sum_{j=1}^d \theta_j^2 \sum_{i=1}^n f_j(x_i)^2 \leqslant \|\theta\|_2^2 \left(\frac{n}{d}+1\right)\frac{L^2}{4\,d^2} \leqslant \left(\frac{n}{d}+1\right)\frac{L^2}{4\,d}\,,$$

where for the first inequality we used that $\|f_j\|_\infty \leqslant L/(2d)$ and that at most $n/d+1$ design points x_i are in the support of a function f_j, while the second inequality follows from $\|\theta\|_2^2 \leqslant d\|\theta\|_\infty^2 \leqslant d$ as $\theta \in \Theta$. Summarizing, we proved

$$\forall\,\theta \in \Theta, \qquad \mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_0) \leqslant \frac{(n+d)L^2}{8\,d^2\sigma^2}\,. \tag{6.53}$$

Second, as in inequality (6.48), we write

$$\begin{aligned}\int_\Theta \mathbb{P}_0\Big(\big\|\widehat{\theta}-\theta\big\|_2^2 \leqslant \rho d\Big)\mathrm{d}\nu(\theta) &= \mathbb{E}_0\Big[\nu\Big(B\big(\widehat{\theta}, \sqrt{\rho d}\big)\cap\Theta\Big)\Big] \\ &\leqslant \frac{\big(\sqrt{\rho d}\big)^d}{2^d}\,\mathrm{Vol}_{\mathbb{R}^d}\big(B(0,1)\big) = \left(\frac{\rho d}{4}\right)^{d/2}\frac{\pi^{d/2}}{\Gamma(1+d/2)}\,,\end{aligned} \tag{6.54}$$

where we used the formula for the volume of the unit Euclidean ball $B(0,1)$ of $\mathbb{R}^d$. As in addition,

$$\Gamma\left(1+\frac{d}{2}\right) = \int_0^{+\infty} t^{d/2}\,\mathrm{e}^{-t}\mathrm{d}t \geqslant \int_{d/2}^{+\infty} (d/2)^{d/2}\,\mathrm{e}^{-t}\mathrm{d}t = \left(\frac{d/2}{\mathrm{e}}\right)^{d/2}\,,$$

we finally get with the choice $\rho = 1/(2\pi\mathrm{e})$

$$\int_\Theta \mathbb{P}_0\Big(\big\|\widehat{\theta}-\theta\big\|_2^2 \leqslant \rho d\Big)\mathrm{d}\nu(\theta) \leqslant \left(\frac{\rho d}{4}\right)^{d/2}\frac{\pi^{d/2}}{\Gamma(1+d/2)} \leqslant \left(\frac{\rho\pi\mathrm{e}}{2}\right)^{d/2} = 4^{-d/2} = 2^{-d} \leqslant \frac{1}{8} \tag{6.55}$$

for $d \geqslant 3$. For $d = 2$, using that $\Gamma(2) = 1$, we see that the final upper bound in (6.54) equals $1/(4\mathrm{e}) \leqslant 1/8 \leqslant 2^{-d}$. We use the 2^{-d} upper bound for the second term in the left-hand side of (6.52) and the $1/8$ upper bound for its first term.

Now, combining (6.50)–(6.52), we proved so far that, for any integer $d \geqslant 2$ to be determined,

$$\sup_{f\in\mathcal{F}_L} \mathbb{E}_f\Big[\big\|\widehat{f}-f\big\|_2^2\Big] \geqslant \frac{L^2}{24\,\pi\mathrm{e}\,d^2}\left(\frac{2}{3} - \frac{(n+d)L^2}{8\,d^3\sigma^2\log(2)}\right)\,. \tag{6.56}$$

In the rest of the proof we choose

$$d = \left\lceil \left(\frac{n\, L^2/\sigma^2}{\log(2)} \right)^{1/3} \right\rceil .$$

We consider sample sizes n such that

$$n \geqslant \max\left\{ \frac{L}{\sigma\sqrt{\log(2)}},\ 64\log(2)\,\frac{\sigma^2}{L^2} \right\} ;$$

the first condition ensures that $d \leqslant n$ while the second condition entails that $d \geqslant 4$. The bound $d \geqslant 2$ was required above, while the bound $d \leqslant n$ simplifies the lower bound (6.56) into

$$\frac{L^2}{24\,\pi\mathrm{e}\,d^2}\left(\frac{2}{3} - \frac{(n+d)L^2}{8\,d^3\sigma^2\log(2)} \right) \geqslant \frac{L^2}{24\,\pi\mathrm{e}\,d^2}\left(\frac{2}{3} - \frac{n\,L^2}{4\,d^3\sigma^2\log(2)} \right).$$

We substitute the value of d therein, using that

$$\left(\frac{n\,L^2/\sigma^2}{\log(2)} \right)^{1/3} \leqslant d \leqslant 1 + \left(\frac{n\,L^2/\sigma^2}{\log(2)} \right)^{1/3} \leqslant \frac{5}{4}\left(\frac{n\,L^2/\sigma^2}{\log(2)} \right)^{1/3},$$

where the last inequality follows from $n \geqslant 64\log(2)\,\sigma^2/L^2$. We get

$$\begin{aligned}
\frac{L^2}{24\,\pi\mathrm{e}\,d^2}\left(\frac{2}{3} - \frac{n\,L^2}{4\,d^3\sigma^2\log(2)} \right) &\geqslant \frac{L^2}{24\,\pi\mathrm{e}\,d^2}\left(\frac{2}{3} - \frac{1}{4} \right) = \frac{L^2}{24\,\pi\mathrm{e}\,d^2}\frac{5}{12} \\
&\geqslant \frac{L^2}{24\,\pi\mathrm{e}}\left(\frac{1}{(5/4)^2}\left(\frac{n\,L^2/\sigma^2}{\log(2)} \right)^{-2/3} \right)\frac{5}{12} \\
&= C\left(\frac{\sigma^2 L}{n} \right)^{2/3} \qquad \text{where} \qquad C = \frac{\big(\log(2)\big)^{2/3}}{90\,\pi\mathrm{e}} .
\end{aligned}$$

A numerical computation shows that this value of C is larger than 0.001, as claimed. Collecting all bounds, the proof is concluded. □

Supremum versus Euclidean norms. Though the general structure of the proof is similar to that of Proposition 6.8.1, we emphasize a technical difference: here, the support Θ of the prior ν is a ball in the supremum norm instead of the Euclidean norm. The reason is that, contrary to Proposition 6.8.1 where the choice of Θ was not constrained (so that we could choose Θ as a Euclidean ball of arbitrary radius), here, we should choose $\Theta \subset [-1,1]^d$ to ensure the inclusion $\mathcal{F}_\Theta \subset \mathcal{F}_L$. Taking a Euclidean ball of radius at most 1 (and ν a uniform prior on this ball) would have led to a choice ρ of the order of $1/d$ for calculations similar to (6.54) and (6.55) to upper bound the integral at hand by a numerical constant smaller than 1. This would result in a lower bound not of the right orders of magnitude in n, L and σ^2. On the contrary, the ball Θ in the sup norm allowed us to choose ρ as a constant and hence get an optimal lower bound.

Extension to Hölder functions. We can easily generalize the proof above to the set of (β, L)–Hölder functions over $[0,1]$ to get a lower bound of the order of $n^{-2\beta/(2\beta+1)}$, where $\beta > 0$. We first recall the definition of such functions. Let $(p, \alpha) \in \mathbb{N} \times (0,1]$ be such that $\beta = p + \alpha$. A function $f : [0,1] \to \mathbb{R}$ is called (β, L)–Hölder if it is p times differentiable and if for all $(x, y) \in [0,1]^2$,

$$\left|f^{(p)}(x) - f^{(p)}(y)\right| \leqslant L|x-y|^{\alpha}\,.$$

We can indeed use the same construction (but without any discretization) as in Tsybakov [2009, Section 2.6] by choosing a function $\varphi : \mathbb{R} \to \mathbb{R}$ that is infinitely differentiable, $(\beta, 1)$–Hölder, and supported on $(0,1)$, and by considering an orthogonal system of the form

$$f_j(x) = \frac{L}{d^{\beta}}\,\varphi\left(d\left(x - \frac{j-1}{d}\right)\right), \qquad j \in \{1, \dots, d\}\,,$$

for a well-chosen d. The proof then follows exactly the same lines as above, with the same functions of the form $f_\theta = \sum_j \theta_j f_j$, for a parameter θ in $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leqslant 1\right\}$.

Note that this proof technique also works in higher dimensions, i.e., for (β, L)–Hölder functions over the m–dimensional cube $[0,1]^m$. A simple adaptation of the above arguments indeed yields a lower bound of the order of $n^{-2\beta/(2\beta+m)}$, in the same spirit as in the lower bound that Györfi et al. [2002, Theorem 3.2] derived in the regression model with random design.

6.8.2 From Bayesian posteriors to point estimators

We recall below a well-known result that indicates how to construct good point estimators from good Bayesian posteriors (Section 6.8.2 below). One theoretical benefit is that this result can be used to convert known minimax lower bounds for point estimation into minimax lower bounds for posterior concentration rates (Section 6.8.2 below). This technique is thus a—less direct—alternative to the method we presented in Section 6.4.1.

The conversion

The following statement is a nonasymptotic variant of Theorem 2.5 by Ghosal et al. [2000] (see also Chapter 12, Proposition 3 by Le Cam, 1986, as well as Section 5.1 by Hoffmann et al., 2015). We consider the same setting as in Section 6.4.1 and assume in particular that the underlying probability measure is given by $P_\theta^{\otimes n}$.

Proposition 6.8.3 (From Bayesian posteriors to point estimators)**.**
Let $n \geqslant 1$, $\delta > 0$, and $\theta \in \Theta$. Let $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n)$ be any estimator satisfying, $P_\theta^{\otimes n}$–almost surely,

$$\mathbb{P}_\pi\Big(\theta' : \ell(\theta', \widehat{\theta}_n) < \varepsilon_n \,\Big|\, X_{1:n}\Big) \geqslant \sup_{\widetilde{\theta} \in \Theta} \mathbb{P}_\pi\Big(\theta' : \ell\big(\theta', \widetilde{\theta}\big) < \varepsilon_n \,\Big|\, X_{1:n}\Big) - \delta\,. \tag{6.57}$$

Then,

$$P_\theta^{\otimes n}\left(\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right) \geqslant P_\theta^{\otimes n}\Big(\ell\big(\widehat{\theta}_n,\theta\big) \geqslant 2\varepsilon_n\Big)\ . \tag{6.58}$$

This result implies that if $\widehat{\theta}_n$ is a center of a ball that almost maximizes the posterior mass—see assumption (6.57)—and if the posterior mass concentrates around θ at a rate $\varepsilon_n' < \varepsilon_n$—so that the left-hand side of (6.58) vanishes by Markov's inequality—then $\widehat{\theta}_n$ is $(2\varepsilon_n)$–close to θ with high probability. Therefore, at least from a theoretical viewpoint, a good posterior distribution can be converted into a good point estimator, by defining $\widehat{\theta}_n$ based on $\mathbb{P}_\pi(\,\cdot\,|\,X_{1:n})$ such that (6.57) holds, i.e., by taking an approximate argument of the supremum. A measurable such $\widehat{\theta}_n$ exists as soon as Θ is a separable topological space and the function $\widetilde{\theta} \mapsto \mathbb{P}_\pi\big(\theta' : \ell(\theta',\widetilde{\theta}\,) < \varepsilon_n \,\big|\, x_{1:n}\big)$ is lower-semicontinuous for $\mathfrak{m}^{\otimes n}$-almost every $x_{1:n} \in \mathcal{X}^n$ (see the end of the proof of Corollary 6.8.4 for more details).

Proof. Denote by $B_\ell(\theta,\varepsilon) \stackrel{\text{def}}{=} \{\theta' \in \Theta : \ell(\theta',\theta) < \varepsilon\}$ the open ℓ–ball of center θ and radius ε. By the triangle inequality we have the following inclusions of events:

$$\begin{aligned}
\Big\{\ell\big(\widehat{\theta}_n,\theta\big) \geqslant 2\varepsilon_n\Big\} &\subseteq \Big\{B_\ell\big(\widehat{\theta}_n,\varepsilon_n\big) \cap B_\ell(\theta,\varepsilon_n) = \varnothing\Big\} \\
&\subseteq \Big\{\mathbb{P}_\pi\big(B_\ell\big(\widehat{\theta}_n,\varepsilon_n\big) \,\big|\, X_{1:n}\big) + \mathbb{P}_\pi\big(B_\ell(\theta,\varepsilon_n) \,\big|\, X_{1:n}\big) \leqslant 1\Big\} \\
&\subseteq \left\{\mathbb{P}_\pi\big(B_\ell(\theta,\varepsilon_n) \,\big|\, X_{1:n}\big) \leqslant \frac{1+\delta}{2}\right\} \qquad (6.59) \\
&= \left\{1 - \mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) < \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right\} \qquad (6.60) \\
&= \left\{\mathbb{P}_\pi\big(\theta' : \ell(\theta',\theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right\},
\end{aligned}$$

where (6.59) follows from the lower bound $\mathbb{P}_\pi\big(B_\ell\big(\widehat{\theta}_n,\varepsilon_n\big) \,\big|\, X_{1:n}\big) \geqslant \mathbb{P}_\pi\big(B_\ell(\theta,\varepsilon_n) \,\big|\, X_{1:n}\big) - \delta$, which holds by assumption (6.57) on $\widehat{\theta}_n$. This concludes the proof. □

Application to posterior concentration lower bounds

We explained above that a good posterior distribution can be converted into a good point estimator. As noted by Ghosal et al. [2000] this conversion can be used the other way around: if we have a lower bound on the minimax rate of estimation, then Proposition 6.8.3 provides a lower bound on the minimax posterior concentration rate, as formalized in the following corollary. Assumption (6.61) below corresponds to an in-probability minimax lower bound; it is for instance a consequence of (6.46) in the standard multivariate Gaussian model with Euclidean loss.

Corollary 6.8.4. *Let $n \geqslant 1$. Consider the setting of Section 6.4.1, with underlying probability measure $P_\theta^{\otimes n}$ when the unknown parameter is θ. Assume that Θ is a separable*

topological space and that $\widetilde{\theta} \mapsto \ell(\theta', \widetilde{\theta})$ *is continuous for all* $\theta' \in \Theta$*. Assume also that for some absolute constant* $c < 1$*, we have*

$$\text{for all estimators } \widehat{\theta}_n, \qquad \inf_{\theta \in \Theta} P_\theta^{\otimes n}\Big(\ell(\widehat{\theta}_n, \theta) < 2\varepsilon_n\Big) \leqslant c\,. \tag{6.61}$$

Then, for all priors π' *on* Θ*,*

$$\inf_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \leqslant \frac{1+c}{2} < 1\,. \tag{6.62}$$

Proof. Let $\delta > 0$ be a parameter that we will later take arbitrarily small. Fix any estimator $\widehat{\theta}_n$ satisfying (6.57) for the prior π', i.e., that almost maximizes the posterior mass on an open ball of radius ε_n. (See the end of the proof for details on why such a measurable $\widehat{\theta}_n$ exists.) Then, Proposition 6.8.3 used for all $\theta \in \Theta$ entails that

$$\sup_{\theta \in \Theta} P_\theta^{\otimes n}\left(\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right) \geqslant \sup_{\theta \in \Theta} P_\theta^{\otimes n}\Big(\ell\big(\widehat{\theta}_n, \theta\big) \geqslant 2\varepsilon_n\Big) \geqslant 1 - c\,,$$

where the last inequality follows from the assumption (6.61). Now we use Markov's inequality to upper bound the left-hand side above and obtain

$$\begin{aligned}\frac{2}{1-\delta} \sup_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big)\Big] &\geqslant \sup_{\theta \in \Theta} P_\theta^{\otimes n}\left(\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) \geqslant \varepsilon_n \,\big|\, X_{1:n}\big) \geqslant \frac{1-\delta}{2}\right) \\ &\geqslant 1 - c\,.\end{aligned}$$

Letting $\delta \to 0$ and dividing both sides by 2 yields

$$1 - \inf_{\theta \in \Theta} \mathbb{E}_\theta\Big[\mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \theta) < \varepsilon_n \,\big|\, X_{1:n}\big)\Big] \geqslant \frac{1-c}{2}\,.$$

Rearranging terms concludes the proof of (6.62). We now address the technical issue mentioned at the beginning of the proof.

Why a measurable $\widehat{\theta}_n$ *exists.* Note that it is possible to choose $\widehat{\theta}_n$ satisfying (6.57) with π' in a mesurable way as soon as Θ is a separable topological space and

$$\psi : \widetilde{\theta} \in \Theta \longmapsto \mathbb{P}_{\pi'}\big(\theta' : \ell(\theta', \widetilde{\theta}\,) < \varepsilon_n \,\big|\, x_{1:n}\big)$$

is lower-semicontinuous for $\mathfrak{m}^{\otimes n}$–almost every $x_{1:n} \in \mathcal{X}^n$, and thus $\mathbb{P}_\theta^{\otimes n}$–almost surely for all $\theta \in \Theta$. The reason is that, in that case, it is possible to equate the supremum of ψ over Θ to a supremum on a countable subset of Θ. Next, and thanks to the continuity assumption on ℓ, we prove that the desired lower-semicontinuity holds true for all $x_{1:n} \in \mathcal{X}^n$ (not just almost all of them).

To that end, we show the lower-semicontinuity at any fixed $\theta^\star \in \Theta$. Consider any sequence $(\widetilde{\theta}_i)_{i \geqslant 1}$ in Θ converging to $\theta^\star$. For all $x_{1:n} \in \mathcal{X}^n$, by Fatou's lemma applied to

the well-defined probability distribution $\mathbb{P}_{\pi'}(\,\cdot\,|\,x_{1:n})$, we have,

$$\begin{aligned}\liminf_{i\to+\infty}\ \mathbb{P}_{\pi'}\Big(\theta' : \ell\big(\theta',\widetilde{\theta}_i\big) < \varepsilon_n \,\Big|\, x_{1:n}\Big) &= \liminf_{i\to+\infty}\ \mathbb{E}_{\pi'}\Big[\mathbb{1}_{\{\ell(\theta',\widetilde{\theta}_i)<\varepsilon_n\}} \,\Big|\, x_{1:n}\Big] \\ &\geqslant \mathbb{E}_{\pi'}\Big[\underbrace{\liminf_{i\to+\infty} \mathbb{1}_{\{\ell(\theta',\widetilde{\theta}_i)<\varepsilon_n\}}}_{=1 \text{ if } \ell(\theta',\theta^\star)<\varepsilon_n} \,\Big|\, x_{1:n}\Big] \qquad (6.63) \\ &\geqslant \mathbb{P}_{\pi'}\Big(\theta' : \ell\big(\theta',\theta^\star\big) < \varepsilon_n \,\Big|\, x_{1:n}\Big)\,,\end{aligned}$$

where in (6.63) we identify that the lim inf equals 1 as soon as $\ell(\theta',\theta^\star) < \varepsilon_n$ by continuity of $\widetilde{\theta} \mapsto \ell\big(\theta',\widetilde{\theta}\big)$ at $\widetilde{\theta} = \theta^\star$. □

6.8.3 Variations on Theorem 6.6.3

The original result by Birgé [2005, Corollary 1] reads, with the notation of Theorem 6.6.3:

$$\min_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{d_N,\ \frac{\overline{K}}{\log(N)}\right\}, \qquad (6.64)$$

where $(d_N)_{N\geqslant 2}$ is a decreasing sequence, defined as follows, based on functions $r_N : [0,1) \to \mathbb{R}$:

$$r_N(a) = \mathrm{kl}\left(a, \frac{1-a}{N-1}\right) - a\log(N) \qquad \text{and} \qquad d_N = \max\big\{a\in[0,1] :\ r_N(a) \leqslant 0\big\}\,.$$

On the other hand, the simplification by Massart [2007, Section 2.3.4] leads to

$$\min_{1\leqslant i\leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{\frac{2\mathrm{e}-1}{2\mathrm{e}},\ \frac{\overline{K}}{\log(N)}\right\}. \qquad (6.65)$$

Before proving these results, we compare them with Theorem 6.6.3. The values of the c_N of Theorem 1, of the d_N of (6.64) and of $(2\mathrm{e}-1)/(2\mathrm{e})$ are given by (values rounded upwards)

$$\frac{2\mathrm{e}-1}{2\mathrm{e}} \approx 0.8161 \qquad \text{and}$$

N	2	3	7	$+\infty$
c_N	0.7587	0.7127	< 0.67	0.63987
d_N	0.7428	0.7009	$< 2/3$	0.63987

The c_N and d_N are thus extremely close. While the c_N are slightly larger than the d_N (with, however, the same limit), they are easier to compute in practice. (See the closed-form expression for r_N below.) Also, the proof of Theorem 6.6.3 is simpler than the proof of Birgé [2005, Corollary 1]: they rely on the proof scheme but the former involves fewer calculations than the latter. Indeed, let us now prove again Birgé [2005, Corollary 1].

Proof of (6.64). We use the notation of the proof of Theorem 6.6.3 and its beginning. We can assume with no loss of generality that $a \geqslant 1/N$, so that, using the definition of a,

$$\widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant a \leqslant \widetilde{p}\,; \tag{6.66}$$

therefore,

$$\mathrm{kl}\big(\widetilde{p},\widetilde{q}\big) \geqslant \mathrm{kl}\big(a,\widetilde{q}\big) \geqslant \mathrm{kl}\left(a,\frac{1-a}{N-1}\right),$$

since by convexity, $p \mapsto \mathrm{kl}(p,q)$ is increasing on $[q,1]$ and $q \mapsto \mathrm{kl}(p,q)$ is decreasing on $[0,p]$. Combining this with $\overline{K} \geqslant \mathrm{kl}\big(\widetilde{p},\widetilde{q}\big)$, one has proved

$$\overline{K} \geqslant \mathrm{kl}\left(a,\frac{1-a}{N-1}\right) = a\log(N) + r_N(a)\,,$$

from which the conclusion follows after studying the variations of $r_N(a)$ in a and N. This last analytical part of the proof is tedious, as

$$\begin{aligned} r_N(a) &= a\log\left(\frac{a}{1-a}\right) + (1-a)\log\left(\frac{1-a}{1-\dfrac{1-a}{N-1}}\right) + \big(a\log(N-1) - a\log(N)\big) \\ &= \big(a\log(a) + (1-2a)\log(a)\big) + a\log\left(\frac{N-1}{N}\right) + (1-a)\log\left(\frac{N-1}{N-2+a}\right), \end{aligned}$$

and we could overcome these heavy calculations in our proof of Theorem 6.6.3.

Proof of (6.65). For $p \geqslant \log(2)$ and all $q \in [0,1]$,

$$\mathrm{kl}(p,q) \geqslant p\log\left(\frac{1}{q}\right) - \log(2) \geqslant p\log\left(\frac{1}{q}\right) - p = p\log\left(\frac{1}{\mathrm{e}q}\right). \tag{6.67}$$

Equation (6.40) is adapted as

$$a \leqslant \widetilde{p} \qquad \text{and} \qquad \widetilde{q} \leqslant \frac{1-a}{N-1} \leqslant \frac{2(1-a)}{N} \leqslant \frac{1}{\mathrm{e}N}$$

where we used respectively, for the last two inequalities, that $1/(N-1) \leqslant 2/N$ for $N \geqslant 2$ and that, with no loss of generality, $a \geqslant (2\mathrm{e}-1)/(2\mathrm{e})$. In particular, $\mathrm{e}\widetilde{q} \leqslant 1/N$. Combining this with $\overline{K} \geqslant \mathrm{kl}\big(\widetilde{p},\widetilde{q}\big)$ and (6.67), we have proved

$$\overline{K} \geqslant \widetilde{p}\log\left(\frac{1}{\mathrm{e}\widetilde{q}}\right) \geqslant a\log(N)\,,$$

which concludes the proof.

6.8.4 Proofs of basic facts about f–divergences (and thus, about Kullback-Leibler divergences)

The results recalled and re-proved in this section were stated in the main body of the chapter (Sections 6.2 and 6.3) for Kullback-Leibler divergences, which are a special case of f–divergences with $f(x) = x \log x$. We restate them in greater generality and to that end, first recall the definition of f–divergences. Note that these f–divergences will be further studied in Section 6.8.5 below, where we show that the reductions and results of Section 6.3 extend in a straightforward manner to arbitrary f–divergences.

Definition of f–divergences and basic properties

The definition of the Kullback-Leibler divergence can be generalized as follows (see Csiszár, 1963, Ali and Silvey, 1966a and Gushchin, 2003 for further details). Let $f : (0, +\infty) \to \mathbb{R}$ be any convex function satisfying $f(1) = 0$. By convexity, we can define

$$f(0) \stackrel{\text{def}}{=} \lim_{t \downarrow 0} f(t) \in \mathbb{R} \cup \{+\infty\} ;$$

the extended function $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ is still convex.

Maximal slope. Note that, for any $x > 0$, the limit

$$\lim_{t \to +\infty} \frac{f(t) - f(x)}{t - x} = \sup_{t > 0} \frac{f(t) - f(x)}{t - x} \in [0, +\infty]$$

exists since (by convexity) the slope $\big(f(t) - f(x)\big)/(t - x)$ is non-decreasing as t increases. Besides, this limit does not depend on x and equals

$$M_f \stackrel{\text{def}}{=} \lim_{t \to +\infty} \frac{f(t)}{t} \in [0, +\infty] ,$$

which thus represents the maximal slope of f. An inequality that we will repeatedly use and that follows from the two equations above with $t = x + y$ is

$$\forall\, x > 0,\ y > 0 , \qquad \frac{f(x + y) - f(x)}{y} \leqslant M_f .$$

Put differently,

$$\forall\, x \geqslant 0,\ y \geqslant 0, \qquad f(x + y) \leqslant f(x) + y\, M_f , \tag{6.68}$$

where the extension to $y = 0$ is immediate and the one to $x = 0$ follows by continuity of f on $(0, +\infty)$, which itself follows from its convexity.

Lebesgue decomposition of measures. We recall that $\ll$ denotes the absolute continuity between measures and we let $\perp$ denote the fact that two measures are singular. For distributions $\mathbb{P}$ and $\mathbb{Q}$ defined on the same measurable space $(\Omega, \mathcal{F})$, the Lebesgue decomposition of $\mathbb{P}$ with respect to $\mathbb{Q}$ is denoted by

$$\mathbb{P} = \mathbb{P}_{\text{ac}} + \mathbb{P}_{\text{sing}}\,, \qquad \text{where} \qquad \mathbb{P}_{\text{ac}} \ll \mathbb{Q} \quad \text{and} \quad \mathbb{P}_{\text{sing}} \perp \mathbb{Q}\,, \tag{6.69}$$

so that $\mathbb{P}_{\text{ac}}$ and $\mathbb{P}_{\text{sing}}$ are both sub-probabilities (positive measures with total mass smaller than or equal to 1) and, by definition,

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \frac{\mathrm{d}\mathbb{P}_{\text{ac}}}{\mathrm{d}\mathbb{Q}}\,.$$

Definition 6.8.5. The f–divergence $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) = \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\text{sing}}(\Omega)\, M_f\,. \tag{6.70}$$

The existence of the integral in the right-hand side follows from the general form of Jensen's inequality stated in Lemma 6.8.12 (Appendix 6.8.6) with $\varphi = f$ and $C = [0, +\infty)$. This inequality, together with (6.68), also indicates that $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) \geqslant 0$. Indeed,

$$\int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q} \geqslant f\left(\int_\Omega \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q}\right) = f\big(\mathbb{P}_{\text{ac}}(\Omega)\big)\,,$$

so that by (6.68),

$$\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) \geqslant f\big(\mathbb{P}_{\text{ac}}(\Omega)\big) + \mathbb{P}_{\text{sing}}(\Omega)\, M_f \geqslant f\big(\mathbb{P}_{\text{ac}}(\Omega) + \mathbb{P}_{\text{sing}}(\Omega)\big) = f(1) = 0\,.$$

Concrete and important examples of f–divergences, such as the Hellinger distance and the χ^2–divergence, are discussed in details below. The Kullback-Leibler divergence corresponds to the function $f : x \mapsto x\,\log(x)$. We have $M_f = +\infty$ for the Kullback-Leibler and χ^2–divergences, while $M_f = 1$ for the Hellinger distance.

The data-processing inequality and two major consequences

Lemma 6.8.6 (Data-processing inequality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}^X$ and $\mathbb{Q}^X$ the laws of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{Div}_f\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})\,.$$

Corollary 6.8.7 (Data-processing inequality with expectations of random variables)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions over the same measurable space $(\Omega, \mathcal{F})$, and let X be any random variable on $(\Omega, \mathcal{F})$ taking values in $[0, 1]$. Denote by $\mathbb{E}_{\mathbb{P}}[X]$ and $\mathbb{E}_{\mathbb{Q}}[X]$ the expectations of X under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then,*

$$\mathrm{div}_f\big(\mathbb{E}_{\mathbb{P}}[X], \mathbb{E}_{\mathbb{Q}}[X]\big) \leqslant \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})\,,$$

where $\mathrm{div}_f(p, q)$ denotes the f–divergence between Bernoulli distributions with respective parameters p and q.

Corollary 6.8.8 (Joint convexity of Div_f). *All f–divergences Div_f are jointly convex, i.e., for all probability distributions $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}_1, \mathbb{Q}_2$ over the same measurable space $(\Omega, \mathcal{F})$, and all $\lambda \in (0,1)$,*

$$\mathrm{Div}_f\Big((1-\lambda)\mathbb{P}_1 + \lambda\mathbb{P}_2,\, (1-\lambda)\mathbb{Q}_1 + \lambda\mathbb{Q}_2\Big) \leqslant (1-\lambda)\,\mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\,\mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)\,.$$

Lemma 6.8.6 and Corollary 6.8.8 are folklore knowledge; we provide here complete and elementary proofs mostly for the sake of self-completeness. These proofs are extracted from Ali and Silvey [1966a, Section 4.2], see also Pardo [2006, Proposition 1.2]. They can be refined: Gray [2011, Lemmas 7.5 and 7.6] establishes (6.71) below and then derives some (stronger) data-processing equality (not inequality). These proof techniques do not seem to be well known; indeed, in the literature many proofs of the elementary properties above for the Kullback-Leibler divergence focus on the discrete case (Cover and Thomas, 2006) or use the duality formula for the Kullback-Leibler divergence (Massart, 2007 or Boucheron et al., 2013, in particular Exercise 4.10 therein).

Proof (of Lemma 6.8.6): We recall that $\mathbb{E}_{\mathbb{Q}}$ denotes the expectation with respect to a measure $\mathbb{Q}$. Let X be a random variable from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$. We write the Lebesgue decomposition (6.69) of $\mathbb{P}$ with respect to $\mathbb{Q}$.

We first show that $(\mathbb{P}_{\mathrm{ac}})^X \ll \mathbb{Q}^X$ and that the Radon-Nikodym derivative of $(\mathbb{P}_{\mathrm{ac}})^X$ with respect to $\mathbb{Q}^X$ equals

$$\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X} = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\,\middle|\, X = \cdot\right] \stackrel{\mathrm{def}}{=} \gamma\,; \tag{6.71}$$

i.e., γ is any measurable function such that $\mathbb{Q}$–almost surely, $\mathbb{E}_{\mathbb{Q}}\big[(\mathrm{d}\mathbb{P}_{\mathrm{ac}}/\mathrm{d}\mathbb{Q})\,\big|X\big] = \gamma(X)$. Indeed, using that $\mathbb{P}_{\mathrm{ac}} \ll \mathbb{Q}$, we have, for all $A \in \mathcal{F}'$,

$$\begin{aligned}(\mathbb{P}_{\mathrm{ac}})^X(A) = \mathbb{P}_{\mathrm{ac}}(X \in A) &= \int_\Omega \mathbb{1}_A(X)\,\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q} = \int_\Omega \mathbb{1}_A(X)\;\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}\,\middle|\, X\right]\mathrm{d}\mathbb{Q} \qquad (6.72)\\ &= \int_\Omega \mathbb{1}_A(X)\,\gamma(X)\mathrm{d}\mathbb{Q} = \int_{\Omega'} \mathbb{1}_A\,\gamma\mathrm{d}\mathbb{Q}^X\,,\end{aligned}$$

where the last equality in (6.72) follows by the tower rule.

Second, by unicity of the Lebesgue decomposition, the decomposition of $\mathbb{P}^X$ with respect to $\mathbb{Q}^X$ is therefore given by

$$\mathbb{P}^X = \big(\mathbb{P}^X\big)_{\mathrm{ac}} + \big(\mathbb{P}^X\big)_{\mathrm{sing}} \qquad \begin{array}{ll}\text{where} & \big(\mathbb{P}^X\big)_{\mathrm{ac}} = (\mathbb{P}_{\mathrm{ac}})^X + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}\\ \text{and} & \big(\mathbb{P}^X\big)_{\mathrm{sing}} = \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}\,.\end{array}$$

The inner $_{\mathrm{ac}}$ and $_{\mathrm{sing}}$ symbols refer to the pair $\mathbb{P}, \mathbb{Q}$ while the outer $_{\mathrm{ac}}$ and $_{\mathrm{sing}}$ symbols refer to $\mathbb{P}^X, \mathbb{Q}^X$.

We use this decomposition for the first equality below and integrate (6.68) for the first inequality below:

$$\begin{aligned}
\mathrm{Div}_f\big(\mathbb{P}^X, \mathbb{Q}^X\big) &= \int_{\Omega'} f\left(\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X} + \frac{\mathrm{d}\big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}^X}\right)\mathrm{d}\mathbb{Q}^X + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}(\Omega')\, M_f \\
&\leqslant \int_{\Omega'} f\left(\frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}})^X}{\mathrm{d}\mathbb{Q}^X}\right)\mathrm{d}\mathbb{Q}^X + \Big(\big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{ac}}(\Omega') + \big(\mathbb{P}_{\mathrm{sing}}\big)^X_{\mathrm{sing}}(\Omega')\Big)\, M_f \\
&= \int_{\Omega'} f(\gamma)\mathrm{d}\mathbb{Q}^X + (\mathbb{P}_{\mathrm{sing}})^X(\Omega')\, M_f \\
&= \int_{\Omega} f\big(\gamma(X)\big)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\, M_f \\
&= \int_{\Omega} f\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\,\middle|X\right]\right)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\, M_f \\
&\leqslant \int_{\Omega} \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\,\middle|X\right]\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\, M_f \qquad (6.73) \\
&= \int_{\Omega} f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q} + \mathbb{P}_{\mathrm{sing}}(\Omega)\, M_f = \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})\,,
\end{aligned}$$

where the inequality in (6.73) is a consequence of the conditional Jensen's inequality in its general form stated in Appendix 6.8.6, Lemma 6.8.13, with $\varphi = f$ and $C = [0, +\infty)$, and where the final equality follows from the tower rule. □

We continue with the proof of Corollary 6.8.7, which is (almost) extracted from Chapter 2, Lemma 2.2.3: it was proved therein for Kullback-Leibler divergences.

Proof (of Corollary 6.8.7): We augment the underlying measurable space into $\Omega \times [0,1]$, where $[0,1]$ is equipped with the Borel σ–algebra $\mathrm{Ber}\big([0,1]\big)$ and the Lebesgue measure $\mathfrak{m}$. We denote by $\mathbb{P} \otimes \mathfrak{m}$ and $\mathbb{Q} \otimes \mathfrak{m}$ the product distributions of $\mathbb{P}$ and $\mathfrak{m}$, $\mathbb{Q}$ and $\mathfrak{m}$. We write the Lebesgue decomposition $\mathbb{P} = \mathbb{P}_{\mathrm{ac}} + \mathbb{P}_{\mathrm{sing}}$ of $\mathbb{P}$ with respect to $\mathbb{Q}$, and deduce from it the Lebesgue decomposition of $\mathbb{P} \otimes \mathfrak{m}$ with respect to $\mathbb{Q} \otimes \mathfrak{m}$: the absolutely continuous part is given by $\mathbb{P}_{\mathrm{ac}} \otimes \mathfrak{m}$, with density

$$(\omega, x) \in \Omega \times [0,1] \longmapsto \frac{\mathrm{d}(\mathbb{P}_{\mathrm{ac}} \otimes \mathfrak{m})}{\mathrm{d}(\mathbb{Q} \otimes \mathfrak{m})}(\omega, x) = \frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(\omega)\,,$$

while the singular part is given by $\mathbb{P}_{\mathrm{sing}} \otimes \mathfrak{m}$, a subprobability with total mass $\mathbb{P}_{\mathrm{sing}}(\Omega)$. In particular,

$$\mathrm{Div}_f\big(\mathbb{P} \otimes \mathfrak{m},\, \mathbb{Q} \otimes \mathfrak{m}\big) = \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})\,.$$

Now, for all events $E \in \mathcal{F} \otimes \mathrm{Ber}\big([0,1]\big)$, the data-processing inequality (Lemma 6.8.6) ensures that

$$\mathrm{Div}_f\big(\mathbb{P} \otimes \mathfrak{m},\, \mathbb{Q} \otimes \mathfrak{m}\big) \geqslant \mathrm{Div}_f\Big((\mathbb{P} \otimes \mathfrak{m})^{\mathbb{1}E},\ (\mathbb{Q} \otimes \mathfrak{m})^{\mathbb{1}E}\Big) = \mathrm{div}_f\big((\mathbb{P} \otimes \mathfrak{m})(E),\, (\mathbb{Q} \otimes \mathfrak{m})(E)\big)\,,$$

where the final equality is by mere definition of div_f as the f–divergence between Bernoulli distributions. The proof is concluded by noting that for the choice of $E = \big\{(\omega, x) \in \Omega \times [0,1] : x \leqslant X(\omega)\big\}$, Tonelli's theorem ensures that

$$(\mathbb{P} \otimes \mathfrak{m})(E) = \int_\Omega \left(\int_{[0,1]} \mathbb{1}_{\{x \leqslant X(\omega)\}} \, \mathrm{d}\mathfrak{m}(x) \right) \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}_{\mathbb{P}}[X] \,,$$

and, similarly, $(\mathbb{Q} \otimes \mathfrak{m})(E) = \mathbb{E}_{\mathbb{Q}}[X]$. □

The joint convexity of Div_f (Corollary 6.8.8) may be proved directly, in two steps. First, the log-sum inequality is generalized into the fact that the mapping $(p, q) \in [0, +\infty)^2 \mapsto q\, f(p/q)$ is jointly convex. Second, a common dominating measure like $\mu = \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{Q}_1 + \mathbb{Q}_2$ is introduced, Radon-Nikodym derivatives p_j and q_j are introduced for the $\mathbb{P}_j$ and $\mathbb{Q}_j$ with respect to μ, and the generalized log-sum inequality is applied pointwise.

We suggest to see instead Corollary 6.8.8 as an elementary consequence of the data-processing inequality.

Proof (of Corollary 6.8.8): We augment the probability space Ω into $\Omega' = \{1, 2\} \times \Omega$ equipped with the σ–algebra $\mathcal{F}'$ generated by the events $A \times B$, where $A \in \big\{\emptyset, \{1\}, \{2\}, \{1, 2\}\big\}$ and $B \in \mathcal{F}$. We define the random pair (J, X) on this space by the projections

$$X : (j, \omega) \in \{1, 2\} \times \Omega \longmapsto \omega \qquad \text{and} \qquad J : (j, \omega) \in \{1, 2\} \times \Omega \longmapsto j \,,$$

and denote by $\mathbb{P}$ the joint distribution of the random pair (J, X) such that $J \sim 1 + \mathrm{Ber}(\lambda)$ and $X | J \sim \mathbb{P}_J$. More formally, $\mathbb{P}$ is the unique probability distribution on $(\Omega', \mathcal{F}')$ such that, for all $(j, B) \in \{1, 2\} \times \mathcal{F}$,

$$\mathbb{P}\big(\{j\} \times B\big) = \big((1 - \lambda)\mathbb{1}_{\{j=1\}} + \lambda \mathbb{1}_{\{j=2\}}\big)\, \mathbb{P}_j(B) \,.$$

Similarly we define the joint probability distribution $\mathbb{Q}$ on $(\Omega', \mathcal{F}')$ using the conditional distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ instead of $\mathbb{P}_1$ and $\mathbb{P}_2$.

The corollary follows directly from the data-processing inequality $\mathrm{Div}_f\big(\mathbb{P}^X, \mathbb{Q}^X\big) \leqslant \mathrm{Div}_f(\mathbb{P}, \mathbb{Q})$, as the laws of X under $\mathbb{P}$ and $\mathbb{Q}$ are respectively given by

$$\mathbb{P}^X = (1 - \lambda)\mathbb{P}_1 + \lambda \mathbb{P}_2 \qquad \text{and} \qquad \mathbb{Q}^X = (1 - \lambda)\mathbb{Q}_1 + \lambda \mathbb{Q}_2 \,,$$

while elementary calculations show that $\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) = (1 - \lambda)\, \mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\, \mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)$.

Indeed, for the latter point, we consider the Lebesgue decompositions of $\mathbb{P}_j$ with respect to $\mathbb{Q}_j$, where $j \in \{1, 2\}$:

$$\mathbb{P}_j = \mathbb{P}_{j,\mathrm{ac}} + \mathbb{P}_{j,\mathrm{sing}} \,, \qquad \text{where} \qquad \mathbb{P}_{j,\mathrm{ac}} \ll \mathbb{Q}_j \quad \text{and} \quad \mathbb{P}_{j,\mathrm{sing}} \perp \mathbb{Q}_j \,.$$

The (unique) Lebesgue decomposition of $\mathbb{P} = \mathbb{P}_{\mathrm{ac}} + \mathbb{P}_{\mathrm{sing}}$ with respect to $\mathbb{Q}$ is then given by

$$\frac{\mathrm{d}\mathbb{P}_{\mathrm{ac}}}{\mathrm{d}\mathbb{Q}}(j, \omega) = \mathbb{1}\{j = 1\}\, \frac{\mathrm{d}\mathbb{P}_{1,\mathrm{ac}}}{\mathrm{d}\mathbb{Q}_1}(\omega) + \mathbb{1}\{j = 2\}\, \frac{\mathrm{d}\mathbb{P}_{2,\mathrm{ac}}}{\mathrm{d}\mathbb{Q}_2}(\omega)$$

and for all $(j, B) \in \{1,2\} \times \mathcal{F}$,

$$\mathbb{P}_{\text{sing}}\big(\{j\} \times B\big) = \big((1-\lambda)\mathbb{1}_{\{j=1\}} + \lambda \mathbb{1}_{\{j=2\}}\big)\, \mathbb{P}_{j,\text{sing}}(B)\,.$$

This entails that

$$\begin{aligned}
\mathrm{Div}_f(\mathbb{P}, \mathbb{Q}) &= \int_{\{1,2\}\times\Omega} f\bigg(\frac{\mathrm{d}\mathbb{P}_{\text{ac}}}{\mathrm{d}\mathbb{Q}}(j,\omega)\bigg)\mathrm{d}\mathbb{Q}(j,\omega) + \mathbb{P}_{\text{sing}}\big(\{1,2\}\times\Omega\big)\, M_f \\
&= (1-\lambda)\int_\Omega f\bigg(\frac{\mathrm{d}\mathbb{P}_{\text{ac}}}{\mathrm{d}\mathbb{Q}}(1,\omega)\bigg)\mathrm{d}\mathbb{Q}_1(\omega) + \lambda\int_\Omega f\bigg(\frac{\mathrm{d}\mathbb{P}_{\text{ac}}}{\mathrm{d}\mathbb{Q}}(2,\omega)\bigg)\mathrm{d}\mathbb{Q}_2(\omega) \\
&\qquad + \big((1-\lambda)\,\mathbb{P}_{1,\text{sing}}(\Omega) + \lambda\,\mathbb{P}_{2,\text{sing}}(\Omega)\big)\, M_f \\
&= (1-\lambda)\,\mathrm{Div}_f(\mathbb{P}_1, \mathbb{Q}_1) + \lambda\,\mathrm{Div}_f(\mathbb{P}_2, \mathbb{Q}_2)\,.
\end{aligned}$$

□

6.8.5 Extensions of the reductions of Section 6.3 to f–divergences

We recall that f–divergences are based on convex functions $f : (0, +\infty) \to \mathbb{R}$ extended at 0 via

$$f(0) \overset{\text{def}}{=} \lim_{t\downarrow 0} f(t) \in \mathbb{R} \cup \{+\infty\}$$

and such that $f(1) = 0$.

Reduction to Bernoulli distributions

We denote by div_f the f–divergence between Bernoulli distributions: for all $(p, q) \in [0,1]^2$,

$$\mathrm{div}_f(p, q) \overset{\text{def}}{=} \mathrm{Div}_f\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big)\,.$$

Because f–divergences are also jointly convex and enjoy a data-processing inequality (see Lemma 6.8.6 and Corollaries 6.8.7 and 6.8.8 in Appendix 6.8.4) the various reductions considered in Section 6.3.1 hold as well. We only illustrate the reduction by considering the simplest one, stated in (6.5), and the most general one, stated in (6.9); with the notation used therein, we have

$$\mathrm{div}_f\bigg(\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i),\, \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)\bigg) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{div}_f\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \frac{1}{N}\sum_{i=1}^N \mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q}_i)\,, \tag{6.74}$$

and (putting aside all measurability issues)

$$\begin{aligned}
\mathrm{div}_f\bigg(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta),\, \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\mathrm{d}\nu(\theta)\bigg) &\leqslant \int_\Theta \mathrm{div}_f\Big(\mathbb{E}_{\mathbb{P}_\theta}\big[Z_\theta\big],\, \mathbb{E}_{\mathbb{Q}_\theta}\big[Z_\theta\big]\Big)\mathrm{d}\nu(\theta) \\
&\leqslant \int_\Theta \mathrm{Div}_f(\mathbb{P}_\theta, \mathbb{Q}_\theta)\mathrm{d}\nu(\theta)\,.
\end{aligned} \tag{6.75}$$

It thus suffices to lower bound div_f to obtain bounds of interest, as we did for kl in Section 6.3.2. We propose below such lower bounds for the χ^2 divergence and the Hellinger distance.

Lower bound on div_f for the χ^2 divergence

This case corresponds to $f(x) = x^2 - 1$. The associated divergence equals, when $\mathbb{P} \ll \mathbb{Q}$,

$$\chi^2(\mathbb{P}, \mathbb{Q}) = \int_\Omega \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)^2 \mathrm{d}\mathbb{Q} - 1\,.$$

A direct calculation indicates that for all $(p, q) \in [0,1]^2$,

$$\chi^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) = \frac{(p-q)^2}{q(1-q)} \geqslant \frac{(p-q)^2}{q}$$

in this case. We get, for instance, the following result based on the reduction (6.74), which corresponds to Proposition 6.2.1 for Kullback-Leibler divergences.

Lemma 6.8.9. *Given an underlying measurable space, for all probability pairs $\mathbb{P}_i$, $\mathbb{Q}_i$ and all events A_i (non necessarily disjoint), where $i \in \{1, \dots, N\}$, with $0 < \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) < 1$, we have*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i) + \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A_i)}\sqrt{\frac{1}{N}\sum_{i=1}^N \chi^2\big(\mathbb{P}_i, \mathbb{Q}_i\big)}\,. \tag{6.76}$$

In particular, if $N \geqslant 2$ and the A_i form a partition,

$$\frac{1}{N}\sum_{i=1}^N \mathbb{P}_i(A_i) \leqslant \frac{1}{N} + \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N}\inf_{\mathbb{Q}}\sum_{i=1}^N \chi^2\big(\mathbb{P}_i, \mathbb{Q}\big)}\,.$$

Lower bound on div_f for the Hellinger distance

This case corresponds to $f(x) = \big(\sqrt{x}-1\big)^2$, for which $M_f = 1$. The associated divergence equals, when $\mathbb{P} \ll \mathbb{Q}$,

$$H^2(\mathbb{P}, \mathbb{Q}) = \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}} - 1\right)^2 \mathrm{d}\mathbb{Q} = 2\left(1 - \int_\Omega \sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}}\,\mathrm{d}\mathbb{Q}\right)$$

and always lies in $[0, 2]$. A direct calculation indicates that for all $p \in [0,1]$ and $q \in (0,1)$,

$$h^2(p,q) \overset{\text{def}}{=} H^2\big(\mathrm{Ber}(p), \mathrm{Ber}(q)\big) = 2\left(1 - \left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right)\right),$$

and further direct calculations in the cases $q = 0$ and $q = 1$ show that this formula remains valid in these cases.

Resorting to the Cauchy-Schwarz inequality. The Cauchy-Schwarz inequality indicates that

$$\sqrt{pq} + \sqrt{(1-q)(1-p)} \leqslant \sqrt{\big(p + (1-q)\big)\big(q + (1-p)\big)} = \sqrt{1 - (p-q)^2}\,,$$

or put differently, that $h^2(p,q) \geqslant 2\Big(1 - \sqrt{1-(p-q)^2}\Big)$, thus

$$p \leqslant q + \sqrt{1 - \big(1 - h^2(p,q)/2\big)^2} = q + \sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}\,, \tag{6.77}$$

which is one of Le Cam's inequalities. This bound is clean and clear enough for the reader to be able to state consequences of it, e.g., in the spirit of Proposition 6.2.1 for Kullback-Leibler divergences or Lemma 6.8.9 for χ^2 divergences. In particular, if the A_i form a partition,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leqslant \frac{1}{N} + \inf_{\mathbb{Q}} \sqrt{\frac{1}{N}\sum_{i=1}^{N} H^2\big(\mathbb{P}_i, \mathbb{Q}\big)}\sqrt{1 - \frac{1}{4N}\sum_{i=1}^{N} H^2\big(\mathbb{P}_i, \mathbb{Q}\big)}\,.$$

Solving for p. This is the path followed by Guntuboyina [2011, Example II.6]; as we prove below, we get

$$p \leqslant q + (1-2q)\,h^2(p,q)\big(1 - h^2(p,q)/4\big) + 2\sqrt{q(1-q)}\,\big(1 - h^2(p,q)/2\big)\sqrt{h^2(p,q)\big(1 - h^2(p,q)/4\big)}\,. \tag{6.78}$$

It can be seen that this bound is a general expression of the bound stated by Guntuboyina [2011, Example II.6]. This bound is slightly tighter than (6.77), by construction (as we solve exactly an equation and perform no bounding) but it is much less readable. It anyway leads to similar conclusions in practice.

Proof. Assuming that $\underline{h}^2 = h^2(p,q)$ is given and fixed, we consider the equation, for the unknown $x \in [0,1]$,

$$\underline{h}^2 = 2\left(1 - \left(\sqrt{q}\sqrt{x} + \sqrt{1-q}\,\sqrt{1-x}\right)\right);$$

this equation is satisfied for $x = p$, by definition of $h^2(p,q)$. Rearranging it, we get the equivalent equation

$$(1-x)(1-q) = \big(1 - \underline{h}^2/2 - \sqrt{q}\sqrt{x}\big)^2 = \big(1 - \underline{h}^2/2\big)^2 - 2\big(1 - \underline{h}^2/2\big)\sqrt{q}\,\sqrt{x} + qx\,,$$

or equivalently again,

$$x - 2\big(1 - \underline{h}^2/2\big)\sqrt{q}\,\sqrt{x} + \big(1 - \underline{h}^2/2\big)^2 - 1 + q = 0\,.$$

Solving this second-order equation for $\sqrt{x}$, we see that all solutions $\sqrt{x}$, including $\sqrt{p}$, are smaller than the largest root; in particular,

$$\sqrt{p} \leqslant \big(1-\underline{h}^2/2\big)\sqrt{q} + \underbrace{\sqrt{\big(1-\underline{h}^2/2\big)^2 q - \big(1-\underline{h}^2/2\big)^2 + 1 - q}}_{=\sqrt{(1-q)\,\underline{h}^2(1-\underline{h}^2/4)}} .$$

Put differently,

$$\begin{aligned} p &\leqslant \big(1-\underline{h}^2/2\big)^2 q + (1-q)\,\underline{h}^2\big(1-\underline{h}^2/4\big) + 2\sqrt{q(1-q)}\,\big(1-\underline{h}^2/2\big)\sqrt{\underline{h}^2\big(1-\underline{h}^2/4\big)} \\ &= q + (1-2q)\,h^2(p,q)\big(1-h^2(p,q)/4\big) \\ &\qquad + 2\sqrt{q(1-q)}\,\big(1-h^2(p,q)/2\big)\sqrt{h^2(p,q)\big(1-h^2(p,q)/4\big)}\,, \end{aligned}$$

which was the expression to obtain. □

Finding a good constant alternative $\mathbb{Q}$

Consider, for example, the reduction based on a convex combination $\alpha = (\alpha_1, \ldots, \alpha_N)$, with all $\alpha_i > 0$:

$$\mathrm{div}_f\left(\sum_{i=1}^N \alpha_i \mathbb{P}_i(A_i), \sum_{i=1}^N \alpha_i\, \mathbb{Q}_i(A_i)\right) \leqslant \sum_{i=1}^N \alpha_i\, \mathrm{div}_f\big(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i)\big) \leqslant \sum_{i=1}^N \alpha_i\, \mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q}_i)\,,$$

which is more general than (6.74) but less general than (6.75).

We wonder, under the constraint that only one fixed alternative distribution $\mathbb{Q}_i = \mathbb{Q}$ is considered, which such alternative to pick. That is, we want to compute or at least upper bound

$$\inf_{\mathbb{Q}} \sum_{i=1}^N \alpha_i\, \mathrm{Div}_f(\mathbb{P}_i, \mathbb{Q})\,;$$

distributions $\mathbb{Q}$ that (approximatively) reach the infimum should be used, at least from a theoretical viewpoint. Sometimes calculations are easier in practice for some specific $\mathbb{Q}$, as we illustrated, for instance, in Section 6.4.2. Otherwise, the lemma below indicates a good candidate, given by the weighted average $\overline{\mathbb{P}}_\alpha$ of the distributions $\mathbb{P}_i$.

To appreciate its performance, we denote by

$$B_f(\alpha) = \max_{j=1,\ldots,N} \mathrm{Div}_f(\delta_j, \alpha)$$

the maximal f–divergence between a Dirac mass δ_j at j and the convex combination α. This bound equals $\log\big(1/\min\{\alpha_1, \ldots, \alpha_N\}\big)$ for a Kullback-Leibler divergence and $1/\min\{\alpha_1, \ldots, \alpha_N\} - 1$ for the χ^2–divergence.

Lemma 6.8.10. *Let $\mathbb{P}_1, \dots, \mathbb{P}_N$ be N probability distributions over the same measurable space $(\Omega, \mathcal{F})$ and let $\alpha = (\alpha_1, \dots, \alpha_N)$ be a convex combination made of positive weights. Then,*

$$\inf_{\mathbb{Q}} \sum_{i=1}^{N} \alpha_i \operatorname{Div}_f(\mathbb{P}_i, \mathbb{Q}) \leqslant \sum_{i=1}^{N} \alpha_i \operatorname{Div}_f\big(\mathbb{P}_i, \overline{\mathbb{P}}_\alpha\big) \leqslant B_f(\alpha)\,,$$

where the infimun is over all probability distributions $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ and where $\overline{\mathbb{P}}_\alpha \overset{\text{def}}{=} \sum_{i=1}^{N} \alpha_i \mathbb{P}_i$.

The first inequality holds with equality in the case of the Kullback-Leibler divergence, as follows from the so-called compensation equality (see, e.g., Yang and Barron, 1999 or Guntuboyina, 2011, Example II.4): assuming with no loss of generality in this case (since $M_f = +\infty$) that $\mathbb{P}_j \ll \mathbb{Q}$ for all $j \in \{1, \dots, N\}$, we have $\overline{\mathbb{P}}_\alpha \ll \mathbb{Q}$ and $\mathrm{d}\mathbb{P}_j/\mathrm{d}\mathbb{Q} = (\mathrm{d}\mathbb{P}_j/\mathrm{d}\overline{\mathbb{P}}_\alpha)(\mathrm{d}\overline{\mathbb{P}}_\alpha/\mathrm{d}\mathbb{Q})$, which entails

$$\sum_{i=1}^{N} \alpha_i \operatorname{KL}(\mathbb{P}_i, \mathbb{Q}) = \sum_{i=1}^{N} \alpha_i \int \left(\log \frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\overline{\mathbb{P}}_\alpha} + \log \frac{\mathrm{d}\overline{\mathbb{P}}_\alpha}{\mathrm{d}\mathbb{Q}} \right) \mathrm{d}\mathbb{P}_i = \left(\sum_{i=1}^{N} \alpha_i \operatorname{KL}\big(\mathbb{P}_i, \overline{\mathbb{P}}_\alpha\big) \right) + \operatorname{KL}\big(\overline{\mathbb{P}}_\alpha, \mathbb{Q}\big)\,,$$

where we used that $\sum_{i=1}^{N} \alpha_i \mathrm{d}\mathbb{P}_i = \mathrm{d}\overline{\mathbb{P}}_\alpha$. So, indeed, the considered infimum is achieved at $\mathbb{Q} = \overline{\mathbb{P}}$.

Proof. The first inequality follows from the choice $\mathbb{Q} = \overline{\mathbb{P}}_\alpha$. For the second inequality, we proceed as in Corollary 6.8.8 and consider the following probability distributions over $\{1, \dots, N\} \times \Omega$: for all $j \in \{1, \dots, N\}$ and all $B \in \mathcal{F}$,

$$\widetilde{\mathbb{P}}\big(\{j\} \times B\big) = \alpha_j\, \mathbb{P}_j(B) \qquad \text{and} \qquad \widetilde{\mathbb{Q}}\big(\{j\} \times B\big) = \alpha_j\, \overline{\mathbb{P}}_\alpha(B)\,.$$

Note that because $\alpha_i > 0$ for all i, we have $\mathbb{P}_j \ll \overline{\mathbb{P}}_\alpha$ for all j. Thus, $\widetilde{\mathbb{P}} \ll \widetilde{\mathbb{Q}}$, with Radon-Nikodym derivative given by

$$(j, \omega) \in \{1, \dots, N\} \times \Omega \;\longmapsto\; \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\widetilde{\mathbb{Q}}}(j, \omega) = \frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) \overset{\text{def}}{=} p_j(\omega)\,.$$

By uniqueness and linearity of the Radon-Nikodym derivatives, we thus have, for $\overline{\mathbb{P}}_\alpha$–almost all ω,

$$\sum_{j=1}^{N} \alpha_j\, p_j(\omega) = \sum_{j=1}^{N} \alpha_j \frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) = \frac{\mathrm{d}\overline{\mathbb{P}}_\alpha}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega) = 1\,, \qquad \text{where} \quad \forall k \in \{1, \dots, N\},\ \alpha_k p_k(\omega) \geqslant 0\,;$$

that is, $\alpha p(\omega) = \big(\alpha_j p_j(\omega)\big)_{1 \leqslant j \leqslant N}$ is a probability distribution over $\{1, \dots, N\}$. (It corresponds to the conditional distribution of j given ω in the probabilistic model $j \sim \alpha$ and $\omega | j \sim \mathbb{P}_j$.)

We now compute $\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big)$ in two different ways. All manipulations below are valid because all integrals defining f–divergences exist (see the comments after the statement of Definition 6.8.5, as well as the first part of the proof of Lemma 6.8.12). Integrating over j first,

$$\begin{aligned}\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big) &= \int_{\{1,\dots,N\}\times\Omega} f\left(\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\widetilde{\mathbb{Q}}}(j,\omega)\right)\mathrm{d}\widetilde{\mathbb{Q}}(j,\omega)\\ &= \sum_{j=1}^N \alpha_j \int_\Omega f\left(\frac{\mathrm{d}\mathbb{P}_j}{\mathrm{d}\overline{\mathbb{P}}_\alpha}(\omega)\right)\mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) = \sum_{j=1}^N \alpha_j\,\mathrm{Div}_f\big(\mathbb{P}_j,\overline{\mathbb{P}}_\alpha\big)\ .\end{aligned}$$

On the other hand, integrating over ω first,

$$\begin{aligned}\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big) &= \int_\Omega \left(\sum_{j=1}^N f\big(p_j(\omega)\big)\,\alpha_j\right)\mathrm{d}\overline{\mathbb{P}}_\alpha(\omega)\\ &= \int_\Omega \left(\sum_{j=1}^N f\left(\frac{\alpha_j p_j(\omega)}{\alpha_j}\right)\alpha_j\right)\mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) = \int_\Omega \mathrm{Div}_f\big(\alpha p(\omega),\alpha\big)\mathrm{d}\overline{\mathbb{P}}_\alpha(\omega) \leqslant B_f(\alpha)\,,\end{aligned}$$

where the last inequality follows by noting that, by joint convexity of Div_f (see Corollary 6.2.3),

$$\mathrm{Div}_f\big(\alpha p(\omega),\alpha\big) \leqslant \sum_{j=1}^n \alpha_j p_j(\omega)\,\mathrm{Div}_f\big(\delta_j,\alpha\big) \leqslant B_f(\alpha)\,.$$

Comparing the two obtained expressions for $\mathrm{Div}_f\big(\widetilde{\mathbb{P}},\widetilde{\mathbb{Q}}\big)$ concludes the proof. □

6.8.6 On Jensen's inequality

Classical statements of Jensen's inequality for convex functions φ on $C \subseteq \mathbb{R}^n$ either assume that the underlying probability measure is supported on a finite number of points or that the convex subset C is open. In the first case, the proof follows directly from the definition of convexity, while in the second case, it is a consequence of the existence of subgradients. In both cases, it is assumed that the function φ under consideration only takes finite values. In this chapter, Jensen's inequality is applied several times to non-open convex sets C, like $C = [0,1]^2$ or $C = [0,+\infty)$ and/or convex functions φ that can possibly be equal to $+\infty$ at some points.

The restriction of C being open is easy to drop when the dimension equals $n = 1$, i.e., when C is an interval; it was dropped, e.g., by Ferguson [1967, pages 74–76] in higher dimensions, thanks to a proof by induction to address possible boundary effects with respect to the arbitrary convex set C. Let $\mathrm{Ber}(\mathbb{R}^n)$ denote the Borel σ–field of $\mathbb{R}^n$.

Lemma 6.8.11 (Jensen's inequality for general convex sets; Ferguson, 1967)**.** *Let $C \subseteq \mathbb{R}^n$ be any non-empty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R}$ be any convex Borel*

function. Then, for all probability measures μ *on* $\big(\mathbb{R}^n, \mathrm{Ber}(\mathbb{R}^n)\big)$ *such that* $\mu(C) = 1$ *and* $\int \|x\| \mathrm{d}\mu(x) < +\infty$*, we have*

$$\int x \mathrm{d}\mu(x) \in C \qquad \textit{and} \qquad \varphi\Big(\int x \mathrm{d}\mu(x)\Big) \leqslant \int_C \varphi(x) \mathrm{d}\mu(x)\,, \tag{6.79}$$

where the integral of φ *against* μ *is well-defined in* $\mathbb{R} \cup \{+\infty\}$*.*

Our contribution is the following natural extension.

Lemma 6.8.12. *The result of Lemma 6.8.11 also holds for any convex Borel function* $\varphi : C \to \mathbb{R} \cup \{+\infty\}$*.*

We rephrase this extension in terms of random variables. Let $C \subseteq \mathbb{R}^n$ be any non-empty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R} \cup \{+\infty\}$ be any convex Borel function. Let X be an integrable random variable from any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\big(\mathbb{R}^n, \mathrm{Ber}(\mathbb{R}^n)\big)$, such that $\mathbb{P}(X \in C) = 1$. Then

$$\mathbb{E}[X] \in C \qquad \text{and} \qquad \varphi\big(\mathbb{E}[X]\big) \leqslant \mathbb{E}\big[\varphi(X)\big]\,,$$

where $\mathbb{E}\big[\varphi(X)\big]$ is well-defined in $\mathbb{R} \cup \{+\infty\}$.

Proof. We first check that $\varphi_- = \max\{-\varphi, 0\}$ is μ–integrable on C, so that the integral of φ against μ is well-defined in $\mathbb{R} \cup \{+\infty\}$. To that end, we will prove that φ is lower bounded on C by an affine function: $\varphi(x) \geqslant a^{\mathrm{T}} x + b$ for all $x \in C$, where $(a, b) \in \mathbb{R}^2$, from which it follows that $\varphi_-(x) \leqslant \|a\|\|x\| + \|b\|$ for all $x \in C$ and thus

$$\int_C \varphi_-(x) \mathrm{d}\mu(x) \leqslant \int_C \Big(\|a\|\|x\| + \|b\|\Big) \mathrm{d}\mu(x) = \|a\| \int_C \|x\| \mathrm{d}\mu(x) + \|b\| < +\infty\,.$$

So, it only remains to prove the affine lower bound. If the domain $\{\varphi < +\infty\}$ is empty, any affine function is suitable. Otherwise, $\{\varphi < +\infty\}$ is a non-empty convex set, so that its relative interior R is also non-empty (see Rockafellar, 1972, Theorem 6.2); we fix $x_0 \in R$. But, by Rockafellar [1972, Theorem 23.4], the function φ admits a subgradient at x_0, that is, there exists $a \in \mathbb{R}^n$ such that $\varphi(x) \geqslant \varphi(x_0) + a^T(x - x_0)$ for all $x \in C$. This concludes the first part of this proof.

In the second part, we show the inequality (6.79) via a reduction to the case of real-valued functions. Indeed, note that if $\mu(\varphi = +\infty) > 0$ then the desired inequality is immediate. We can thus assume that $\mu(\varphi < +\infty) = 1$. But, using Lemma 6.8.11 with the non-empty convex Borel subset $\widetilde{C} = \{\varphi < +\infty\}$ and the real-valued convex Borel function $\widetilde{\varphi} : \widetilde{C} \to \mathbb{R}$ defined by $\widetilde{\varphi}(x) = \varphi(x)$, we get, since $\mu(\widetilde{C}) = 1$:

$$\int x \mathrm{d}\mu(x) \in \widetilde{C} \qquad \text{and} \qquad \widetilde{\varphi}\Big(\int x \mathrm{d}\mu(x)\Big) \leqslant \int_{\widetilde{C}} \widetilde{\varphi}(x) \mathrm{d}\mu(x)\,.$$

Using the facts that $\widetilde{\varphi}(x) = \varphi(x)$ for all $x \in \widetilde{C}$ and that $\mu\big(C \setminus \widetilde{C}\big) = 1 - 1 = 0$ entails (6.79). □

We now complete our extension by tacking the conditional form of Jensen's inequality.

Lemma 6.8.13 (A general conditional Jensen's inequality)**.** *Let $C \subseteq \mathbb{R}^n$ be any non-empty convex Borel subset of $\mathbb{R}^n$ and $\varphi : C \to \mathbb{R} \cup \{+\infty\}$ be any convex Borel function. Let X be an integrable random variable from any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\big(\mathbb{R}^n, \mathrm{Ber}(\mathbb{R}^n)\big)$, such that $\mathbb{P}(X \in C) = 1$. Then, for every sub-σ–field $\mathcal{G}$ of $\mathcal{F}$, we have, $\mathbb{P}$–almost surely,*

$$\mathbb{E}[X \,|\, \mathcal{G}] \in C \qquad \text{and} \qquad \varphi\big(\mathbb{E}[X \,|\, \mathcal{G}]\big) \leqslant \mathbb{E}\big[\varphi(X) \,|\, \mathcal{G}\big]\,,$$

where $\mathbb{E}\big[\varphi(X) \,|\, \mathcal{G}\big]$ is $\mathbb{P}$–almost-surely well-defined in $\mathbb{R} \cup \{+\infty\}$.

Proof. The proof follows directly from the unconditional Jensen's inequality (Lemma 6.8.12 above) and from the existence of regular conditional distributions. More precisely, by Durrett [2010, Theorems 2.1.15 and 5.1.9] applied to the case where $(S, \mathcal{S}) = \big(\mathbb{R}^n, \mathrm{Ber}(\mathbb{R}^n)\big)$, there exists a regular conditional distribution of X given $\mathcal{G}$. That is, there exists a function $K : \Omega \times \mathrm{Ber}(\mathbb{R}^n) \to [0, 1]$ such that:

(P1) for every $B \in \mathrm{Ber}(\mathbb{R}^n)$, $\omega \in \Omega \mapsto K(\omega, B)$ is $\mathcal{G}$–measurable and $\mathbb{P}(X \in B \,\big|\, \mathcal{G}) = K(\,\cdot\,, B)$ $\mathbb{P}$–a.s.;

(P2) for $\mathbb{P}$–almost all $\omega \in \Omega$, the mapping $B \mapsto K(\omega, B)$ is a probability measure over $\big(\mathbb{R}^n, \mathrm{Ber}(\mathbb{R}^n)\big)$.

Moreover, as a consequence of (P1),

(P1') for every Borel function $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(X)$ is $\mathbb{P}$–integrable or such that g is nonnegative,

$$\int g(x)\, K(\,\cdot\,, \mathrm{d}x) = \mathbb{E}\big[g(X) \,\big|\, \mathcal{G}\big] \qquad \mathbb{P}\text{–a.s.}$$

Now, given our assumptions and thanks to (P1) and (P1'):

(P3) by $\mathbb{P}(X \in C) = 1$ we also have $K(\,\cdot\,, C) = \mathbb{P}(X \in C \,|\, \mathcal{G}) = 1$ $\mathbb{P}$–a.s.;

(P4) since X is $\mathbb{P}$–integrable, so is $\int \|x\|\, K(\,\cdot\,, \mathrm{d}x) = \mathbb{E}\big[\|X\| \,\big|\, \mathcal{G}\big]$, which is therefore $\mathbb{P}$–a.s. finite.

We apply Lemma 6.8.12 with the probability measures $\mu_\omega = K(\omega, \,\cdot\,)$, for those ω for which the properties stated in (P2), (P3) and (P4) actually hold; these ω are $\mathbb{P}$–almost all elements of Ω. We get, for these ω,

$$\int x\, K(\omega, \mathrm{d}x) \in C \qquad \text{and} \qquad \varphi\bigg(\int x\, K(\omega, \mathrm{d}x)\bigg) \leqslant \int_C \varphi(x)\, K(\omega, \mathrm{d}x)\,,$$

where the integral in the right-hand side is well defined in $\mathbb{R} \cup \{+\infty\}$. Thanks to (P1'), and by decomposing $\varphi(X)$ into $\varphi_-(X)$, which is integrable (see the beginning of the proof of Lemma 6.8.12), and $\varphi_+(X)$, which is nonnegative, we thus have proved that $\mathbb{P}$–a.s.,

$$\mathbb{E}\big[X \,\big|\, \mathcal{G}\big] \in C \qquad \text{and} \qquad \varphi\Big(\mathbb{E}\big[X \,\big|\, \mathcal{G}\big]\Big) \leqslant \mathbb{E}\big[\varphi(X) \big| \mathcal{G}\big]\,,$$

which concludes the proof. □

Bibliography

R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

S. Ali and S. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B. Methodological*, 28: 131–142, 1966a.

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. Series B. Methodological*, 28: 131–142, 1966b.

A. Antos, V. Grover, and C. Szepesvári. Active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 287–302. Springer, 2008.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory* (COLT), COLT'09, pages 217–226. 2009.

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1):55–65, 2010.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases-Volume 8724*, pages 115–131. Springer-Verlag New York, Inc., 2014.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions*. John Wiley and Sons, 1973.

D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

C. Berge. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation, 1963.

L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1615, 2005.

J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, 2013.

S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. *Lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004, 2003.

J. Bretagnolle and C. Huber. Estimation des densités : risque minimax. *Séminaire de Probabilités de Strasbourg*, 12:342–363, 1978.

J. Bretagnolle and C. Huber. Estimation des densités : risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

S. Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, France, 2010.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck and C.-Y. Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pages 638–646, 2013.

S. Bubeck and A. Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Annual Conference on Learning Theory* (COLT), *JMLR W&CP*, volume 30, pages 122–134. 2013a.

S. Bubeck, V. Perchet, and P. Rigollet. Erratum to Bubeck et al. [2013a], 2013b. URL `http://research.microsoft.com/en-us/um/people/sebubeck/pub.html`. "The proof of Theorem 8 is not correct. We do not know if the theorem holds true.".

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

C. Calabro. *The Exponential Complexity of Satisfiability Problems*. PhD thesis, University of California, San Diego, 2009.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3): 1516–1541, 2013.

A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604, 2016.

R. Cerf and P. Petit. A short proof of Cramér's theorem in R. *The American Mathematical Monthly*, 118(10):925–931, 2011.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. Helmbold, R. Schapire, and M. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label-efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2005.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 379–387. Curran Associates, Inc., 2014.

X. Chen, A. Guntuboyina, and Y. Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(219):1–58, 2016.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

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.

R. Combes and A. Proutière. Unimodal bandits without smoothness, 2014. arXiv:1406.7447.

T. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, second edition, 2006.

W. Cowan and M. Katehakis. Asymptotically optimal sequential experimentation under generalized ranking, 2015. arXiv:1510.02041.

H. Cramér. Sur un nouveau théorème limite de la théorie des probabilités. *Actualites Scientifiques et Industrielles*, 736:5–23, 1938.

I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8:85–108, 1963.

I. Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

I. Csiszár and F. Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of the 2016 International Conference on Machine Learning*, ICML'16, pages 1587–1595, 2016.

J. Duchi. *Lecture Notes for Statistics 311/Electrical Engineering 377*, chapter 4, pages 41–48. 2014. URL https://web.stanford.edu/class/stats311/Lectures/lec-05.pdf.

J. Duchi and M. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. 2013. arXiv:1311.2669.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010.

E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

M. Faure, P. Gaillard, B. Gaujal, and V. Perchet. Online learning and game theory. a quick overview with recent results and applications. *ESAIM: Proceedings and Surveys*, 51:246–271, October 2015.

T. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London, 1967.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

M. Frisén. Unimodal regression. *The Statistician*, pages 479–485, 1986.

A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

A. Garivier, E. Kaufmann, and T. Lattimore. On explore-then-commit strategies. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29* (NIPS 2016), pages 784–792. Curran Associates, Inc., 2016.

Z. Geng and N.-Z. Shi. Algorithm as 257: isotonic regression for umbrella orderings. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(3):397–402, 1990.

M. C. Genovese, P. Durez, H. B. Richards, J. Supronik, E. Dokoupilova, V. Mazurov, J. A. Aelion, S.-H. Lee, C. E. Codding, H. Kellner, T. Ikawa, S. Hugot, and S. Mpofu. Efficacy and safety of secukinumab in patients with rheumatoid arthritis: a phase ii, dose-finding, double-blind, randomised, placebo controlled study. *Annals of the Rheumatic Diseases*, 72(6):863–869, 2013.

S. Ghosal, J. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

R. Gray. *Entropy and Information Theory*. Springer, second edition, 2011.

A. Guntuboyina. Lower bounds for the minimax risk using-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

A. Gushchin. On fanos lemma and similar inequalities for the minimax risk. *Probability Theory and Mathematical Statistics*, 67:26–37, 2003.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

T. Han and S. Verdú. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994.

H. Harari-Kermadec. *Vraisemblance empirique généralisée et estimation semi-paramétrique*. PhD thesis, ENSAE ParisTech, 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. *Annals of Statistics*, 43(5):2259–2295, 2015.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16(Dec):3721–3756, 2015.

A. Hoorfar and M. Hassani. Inequalities on the Lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):Article 51, 2008a.

A. Hoorfar and M. Hassani. Inequalities on the Lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):Article 51, 2008b.

X. Hu. Maximum-likelihood estimation under bound restriction and order and uniform bound restrictions. *Statistics & probability letters*, 35(2):165–171, 1997.

I. Ibragimov and R. Has'minskii. *Statistical Estimation: Asymptotic Theory*, volume 16. Springer-Verlag New York, 1981.

I. Ibragimov and R. Has'minskii. Bounds for the risks of non-parametric regression estimates. *Theory of Probability and its Applications*, 27(1):84–99, 1982.

I. Ibragimov and R. Has'minskii. Asymptotic bounds on the quality of the nonparametric regression estimation in L_p. *Journal of Mathematical Sciences*, 24(5):540–550, 1984.

M. Iltis. Sharp asymptotics of large deviations in d. *Journal of Theoretical Probability*, 8(3):501–522, 1995.

C. Jiang. *Online Advertisements and Multi-Armed Bandits*. PhD thesis, University of Illinois at Urbana-Champaign, USA, 2015.

E. Kaufmann. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016.

E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pages 592–600, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1): 1–42, 2016.

M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (UAI'98)*, pages 311–319, 1998.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

S. Kulkarni and G. Lugosi. Minimax lower bounds for the two-armed bandit problem. *IEEE Transactions on Automatic Control*, 45:711–714, 2000.

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(229):1–32, 2016.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

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

T. Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

T. Lattimore. Regret analysis of the anytime optimally confident UCB algorithm. arXiv:1603.08661, 2016.

T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Advances in Neural Information Processing Systems*, pages 1583–1592, 2017.

T. Lattimore. Refining the confidence level for optimistic bandit strategies. 2018. submitted.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

L. Le Cam and G. Yang. *Asymptotics in statistics: some basic concepts*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2000.

C. Le Tourneau, J. J. Lee, and L. L. Siu. Dose escalation methods in phase i cancer clinical trials. *JNCI: Journal of the National Cancer Institute*, 101(10):708–720, 2009.

E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.

A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1690–1698, 2016.

S. Magureanu, R. Combes, and A. Proutière. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 975–999, 2014.

O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514, 2011.

S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004a.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004b.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Procedings of the 2017 Algorithmic Learning Theory Conference*, ALT'17, 2017.

R. Munos et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–129, 2014.

R. Mureika, T. Turner, and P. Wollan. An algorithm for unimodal isotonic regression, with application to locating a maximum, univ. new brunswichk dept. math. Technical report, and Stat. Tech. Report 92–4, 1992.

E. Ordentlich and M. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, 2005.

A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.

C. Pandit and S. Meyn. Worst-case large-deviation asymptotics with application to queueing and information theory. *Stochastic processes and their applications*, 116(5): 724–756, 2006.

L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman & Hall/CRC, 2006.

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. John Wiley and Sons, 1988.

R. Rockafellar. *Convex Analysis*. Princeton University Press, second edition, 1972.

D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

M. Simchowitz, K. Jamieson, and B. Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. *arXiv preprint arXiv:1702.05186*, 2017.

G. Stoltz. An introduction to the prediction of individual sequences: (1) oracle inequalities; (2) prediction with partial monitoring, 2007. Statistics seminar of Université Paris VI and Paris VII, Chevaleret, November 12 and 26, 2007; written version of the pair of seminar talks available upon request.

Q. F. Stout. Optimal algorithms for unimodal regression. *Ann Arbor*, 1001:48109–2122, 2000.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. Technical report, Hewlett-Packard, 2003.

Y. Wu, A. György, and C. Szepesvari. Online learning with Gaussian payoffs and side observations. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28* (NIPS 2015), pages 1360–1368. Curran Associates, Inc., 2015.

A. Xu and M. Raginsky. Information-theoretic lower bounds on Bayes risk in decentralized estimation. 2016. arXiv:1607.00550.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. New York, NY, 1997.