On some adaptivity question in stochastic multi-armed bandits problems

Hédi Hadiji

université
PARIS-SACLAY

On some adaptivity questions in stochastic multi-armed bandits

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques appliquées
Unité de recherche : - Laboratoire de mathématiques d'Orsay
(Faculté des sciences d'Orsay), UMR 8628 CNRS
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 4 décembre 2020, par

Hédi HADIJI

Composition du jury

Gilles Blanchard Professeur, Université Paris-Saclay	Examinateur
Alexandra Carpentier Professeure, Otto von Guericke Universität	Rapporteur
Anatoli Iouditski Professeur, Université Grenoble Alpes	Rapporteur
Vianney Perchet Professeur, CREST, ENSAE	Examinateur
Tim van Erven Associate Professor, Universiteit van Amsterdam	Examinateur
Gilles Stoltz Directeur de recherche, CNRS/Université Paris-Saclay	Directeur
Pascal Massart Professeur, Université Paris-Saclay	Codirecteur

Thèse de doctorat
NNT: 2020UPASM021


université
PARIS-SACLAY
ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)


MATHEMATIQUES
ORSAY


Fondation mathématique
FMJH
Jacques Hadamard

[...] un labyrinthe caché dans un rébus, avec une énigme pour solution ? Oui, cela peut être intéressant, s'il y a le trésor habituel à la clé...
... et même s'il n'y a rien... il y a les arcanes, le mystère, l'ambiguïté, le sphinx, l'allégorie, la charade...
...ce qui compte... c'est le symbole, le jeu, l'aventure, Corto.

(Hugo Pratt, *Mū, la cité perdue*)

Contents

Remerciements

Merci bien évidemment à mes directeurs de thèse, Gilles et Pascal qui m'ont guidé pendant ces années de thèse. Gilles, tu m'as proposé des sujets passionnants, et tu m'as accompagné avec tes exigences de rigueur et de clarté, ta bonne humeur, ta bienveillance et ta disponibilité inégalée. J'espère que mon sens de l'organisation, moins affuté que le tien, ne t'a pas trop donné de sueurs froides. Pascal, peut-être à ton insu, tu m'as donné une grande leçon de persévérance en m'incitant au bon moment à continuer à gratter. Tu m'as suivi de plus loin mais cette thèse n'aurait été la même sans tes précieux conseils.

Merci beaucoup à Alexandra Carpentier et Anatoli Iouditski qui m'ont fait l'honneur de rapporter ma thèse ; merci pour votre lecture attentive et pour vos commentaires. Un grand merci également à Gilles Blanchard, Tim van Erven et Vianney Perchet d'avoir accepté de faire partie de mon jury de thèse. Tim, thanks again for being in my jury, I am quite excited about the work we have started, and I hope there is going to be a lot more, together with Dirk, Sarah and Jack.

Merci aussi à ceux avec qui j'ai eu la chance de collaborer : Aurélien Garivier, Sébastien Gerchinovitz, Jean-Michel Loubes et Pierre Ménard. On dirait bien que j'ai eu un certain tropisme pour Toulouse durant cette thèse. Merci pour les accueils, j'espère y revenir bientôt, et rendre visite à ceux qui sont ailleurs.

Merci aux membres du LMO et en particulier à l'équipe probas/stats, qui a fourni un cadre riche et stimulant à ces années de recherche, dans des conditions matérielles fantastiques. Merci en particulier à Stéphane Nonnenmacher pour son suivi minutieux.

Merci aux camarades doctorants du labo d'Orsay, pour les petites balades en forêt, les courtes pauses café, les discussions échecs, les déjeuners passionnants et les grands voyages en RER : Armand, Margaux, Malo, Hugo, Ernesto, Solenne, Benjamin et tous les autre.

Je remercie bien sûr du fond du coeur mes ami.e.s, grâce à qui j'ai pu oublier les (rares) preuves fausses que j'ai commises. Parmi eux, quelques futures (pas si futurs) brillants chercheurs, quelques futures (quand même) milliardaires, quelques vagabonds, mais tous.tes des *oiseaux de passage*.

> *Oh ! les gens bienheureux !... Tout à coup, dans l'espace,*
> *Si haut qu'il semble aller lentement, un grand vol*
> *En forme de triangle arrive, plane et passe.*
> *Où vont-ils ? Qui sont-ils ? Comme ils sont loin du sol !*

Citons en quelques-un.es (attention j'aime pas choisir, c'est `np.random.permutation` qui s'en occupe) Eric, Ronan, Pierre-Luc, Tom, Maxime, Flora, Jonathan, Antoine, Sam, Alexandre, et tant d'autres, continuez à voltiger à droite à gauche (mais plus à gauche quand même) je compte bien voler à vos côtés.

Merci au bureau de Ronan et Hadrien à l'INRIA, super squat, super ambiance. A Tom et Pierre qui m'ont hébergé au bord de la mer pendant la rédaction du manuscrit, un merci intéressé : on refait ça quand vous voulez.

Quelques mercis en vrac à la BNF, au café Arobase, à Arkose, à Atout-Livre, aux cafetières Bialetti, à Lichess, à Karine des ateliers beaux-arts, à M. Humbert de l'EABJM, et à Yves Duval.

C'est à ma famille que je dois tout. A mes parents, qui chacun.e à leur manière, m'inspirent un peu tous les jours. J'admire votre courage, et je vous remercie pour votre amour. J'embrasse aussi mes grands-mères 多谢. شكر

Merci à Sophie, pour ton soutien, pour tous ces beaux moments.

Chapitre 1.

Vue d'ensemble des résultats

Contenu

1.1. Problèmes de bandits

1.1.1. Généralités sur les bandits à K bras

Protocole d'observation, problème, stratégie. Soit K un entier naturel. Un problème de bandits stochastiques à K bras est une famille de K distributions de probabilités sur $\mathbb{R}$, notée $\underline{\nu} = (\nu_1, \dots, \nu_K)$. On note $[K] = \{1, \dots, K\}$ l'ensemble des entiers entre 1 et K. Pour $a \in [K]$, la distribution ν_a modélise la loi des paiements associés à l'*action* a, aussi appelée le *bras*.

Le protocole qui régit les observations est le suivant. A chaque temps $t \in \{1, 2, \dots, T, \dots, \}$, le statisticien, parfois appelé le joueur, choisit une action $A_t \in [K]$. Il reçoit alors un paiement Y_t, de loi ν_{A_t} conditionnellement à A_t, et tiré indépendamment des choix et des observations précédentes, conditionnellement à A_t.

Un algorithme (ou une stratégie) déterministe est une suite de fonctions $\psi = (\psi_1, \dots, \psi_t, \dots)$, où ψ_t associe le choix A_t aux observations disponibles au temps t, c'est-à-dire à la famille $(A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$. Pour prendre en compte les stratégies aléatoires, on autorise ψ_t à être

aussi fonction d'une variable aléatoire U_t, indépendante des observations précédentes et, par exemple, de loi uniforme sur $[0,1]$. "

Exemples d'application. On imagine la situation suivante. Dans un casino, un joueur fait face à des machines à sous, et peut sélectionner, tour à tour, la machine de son choix. On dit que l'ensemble des machines forme un bandit à plusieurs bras, et que le joueur choisit à chaque tour t un bras A_t, dont le paiement est modélisé par Y_t de loi ν_{A_t}. C'est de cet exemple fantaisiste qu'est tiré le nom de "bandits" ; le bandit à plusieurs bras s'opposant au bandit manchot, capable de dévaliser le pauvre joueur avec un seul bras.

Historiquement, la théorie des bandits s'est inspirée d'applications plus sérieuses, et en particulier des essais cliniques. Dans l'article où apparaît la première formulation d'un problème de bandits, Thompson [1933], l'auteur pose le problème suivant. Des médecins souhaitent traiter une maladie, et disposent pour cela de K médicaments candidats, dont ils ne connaissent pas les efficacités respectives. Des participants à un essai clinique arrivent un par un. A chaque patient, ils proposent un des K médicaments, après quoi ils observent (avant de traiter les patients suivants) l'effet du médicament, c'est-à-dire s'il y a eu ou non guérison.

Si l'on suppose que chaque patient a une probabilité de guérir ne dépendant que du médicament administré, alors ce problème s'ancre bien dans le protocole décrit au paragraphe précédent.

Notations. On supposera toujours *a minima* que les lois des paiements possèdent un premier moment. Introduisons alors quelques notations standards. On note $\mu_a = \mathrm{E}(\nu_a)$ le paiement moyen du bras a, et $\mu^\star = \max\{\mu_a \ : \ a \in [K]\}$ le meilleur paiement moyen. Lorsque cela peut se faire sans ambiguïté, on note $a^\star$ un bras ayant un paiement moyen maximal (on parle de bras optimal). Pour $a \in [K]$, on note $\Delta_a = \mu^\star - \mu_a$ l'écart de sous-optimalité du bras a.

Regret. Dans le problème de bandits standard, l'objectif du statisticien sera alors de maximiser son paiement cumulé. De manière équivalente, celui-ci souhaite minimiser son regret, qui est défini comme la différence entre le paiement moyen obtenu s'il avait joué la meilleure action tout du long, soit $T\mu^\star$, et l'espérance du vrai paiement cumulé obtenu :

$$R_T(\underline{\nu}, \psi) = T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] .$$

On omettra souvent la dépendence en $\underline{\nu}$ ou ψ du regret. Par définition du protocole, le paiement reçu au temps t vérifie $\mathbb{E}[Y_t \mid A_t] = \mu_{A_t}$. Cette identité permet une réécriture éclairante du regret. Notons

$$N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{A_t = a\}}$$

le nombre de fois où l'action $a \in [K]$ a été sélectionnée au temps T. Alors, en conditionnant par A_t, et puisqu'Y_t est de loi ν_{A_t}, on a $\mathbb{E}[Y_t | A_t] = \mu_{A_t}$. Par conséquent, le regret admet la réécriture suivante :

$$\begin{aligned} R_T(\underline{\nu}, \psi) = T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] &= T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{E}\big[Y_t \,\big|\, A_t\big]\Big] \\ &= T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^{T} \mu_{A_t}\Big] = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t}\big] = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(t)] \,. \quad (1.1) \end{aligned}$$

S'intéresser au regret est donc équivalent à étudier le nombre de tirages des bras sous-optimaux.

Comme souvent en statistiques, les résultats théoriques en bandits peuvent être séparés en deux familles. D'un côté, les algorithmes et leurs analyses donnent des bornes supérieures sur le regret face à un problème ; ce sont des résultats de la forme $R_T \leqslant (\dots)$. De l'autre côté, les bornes inférieures mettent en évidence les limites fondamentales de la minimisation du regret et permettent de s'intéresser aux propriétés d'optimalité des algorithmes.

Information préalable et environnement. Avant d'aborder des exemples concrets de stratégies, il nous faut spécifier l'information préalable dont dispose le statisticien sur le problème auquel il fait face. Ainsi, on supposera que le joueur sait que le problème $\underline{\nu}$ contre lequel il joue appartient à un ensemble de problèmes donné, qu'on appellera une classe d'environnements, et qu'on notera $\mathcal{E}$. La classe $\mathcal{E}$ est donc un sous-ensemble de l'ensemble des K–familles de probabilités sur $\mathbb{R}$ possédant un premier moment.

Exemple 1.1 (Bandits gaussiens de variance 1)**.** *Si l'on suppose que les paiements associés à chaque bras sont gaussiens de variance* 1*, alors la classe d'environnements en question est*

$$\mathcal{E} = \big\{(\nu_1, \dots, \nu_K) : \textit{pour tout } a \in [K]\,, \quad \nu_a \textit{ est une distribution gaussienne } \mathcal{N}(\mu_a, 1)\big\}$$

Le cas où les paiements sont bornés est un cadre très souvent étudié, notamment dans l'article fondateur d'Auer et al. [2002a].

Exemple 1.2 (Paiements bornés dans $[0,1]$)**.** *Lorsque l'on suppose que le paiement de chaque bras peut suivre n'importe quelle loi, mais qu'il est borné dans* $[0,1]$*, la classe d'environnements est*

$$\mathcal{E}_{[0,1]} = \big\{(\nu_1, \dots, \nu_K) \ : \ \forall a \in [K]\,, \quad \nu_a\big([0,1]\big) = 1\big\}\,.$$

Les exemples précédents sont des cas particuliers de classes non-structurées : la donnée du paiement d'un bras n'informe en rien sur les paiements des autres bras, et tous les bras jouent un rôle équivalent. Dans le cas non-structuré, la classe s'écrit comme une puissance K–ième (pour le produit cartésien) d'un ensemble $\mathcal{D}$ de mesures de probabilité sur $\mathbb{R}$ — dans notre second exemple, il s'agit de l'ensemble des mesures de probabilité à support dans $[0,1]$. On dira alors que $\mathcal{D}$ est un modèle. Le cas non-structuré est le cadre standard pour les problèmes de bandits.

La donnée de la classe d'environnement $\mathcal{E}$ joue un rôle crucial dans l'étude théorique des problèmes de bandits. Cette donnée intervient bien sûr dans les hypothèses que l'on fait sur le problème auquel le joueur fait face. Elle est surtout particulièrement importante lorsque l'on s'intéresse aux bornes inférieures.

L'objectif de cette thèse est d'appréhender, à travers l'étude d'exemples, le vaste problème de la connaissance imparfaite de l'environnement au joueur. Autrement dit, on souhaite répondre à la question, "Que faire lorsque les informations dont dispose le joueur sont incertaines ?" On reformulera cette question de façon de plus en plus précise au cours de cette introduction.

Avant de commencer à y répondre, présentons les types de garanties auxquelles on peut s'attendre, en nous appuyant sur l'exemple de l'algorithme `UCB`.

1.1.2. Algorithmes et garanties : l'exemple d'`UCB`

Upper Confidence Bounds (`UCB`)

L'algorithme `UCB` est, depuis son analyse moderne dans Auer et al. [2002a], l'algorithme par excellence dans les problèmes de bandits. La simplicité de sa formulation et de son analyse en font une brique fondamentale pour de nombreuses approches. La stratégie est la suivante : au

cours des K premiers tours, chaque bras est tiré une fois, après quoi on associe à chaque bras $a \in [K]$ un indice,

$$U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2\ln T}{N_a(t)}}\,, \quad \text{où} \quad \widehat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s = a\}}\,.$$

La quantité $\widehat{\mu}_a(t)$ est la moyenne empirique des paiements associés au bras a. Le joueur tire alors à chaque tour un bras ayant le plus grand indice

$$A_{t+1} \in \operatorname*{argmax}_{a \in [K]} U_a(t)\,,$$

et un choix arbitraire est fait en cas d'égalité. Par cohérence avec la présentation des résultats ultérieurs, on a décrit ici une version dite "*non-anytime*" d'`UCB`, c'est-à-dire qui utilise la connaissance de l'horizon de temps T, temps auquel on évaluera l'algorithme.

`UCB` appartient à la famille des stratégies à indice : l'indice d'un bras est une quantité calculée exclusivement en fonction des paiements associés à ce bras, et la stratégie sélectionne à chaque tour un bras d'indice maximal. Pour `UCB`, l'indice correspond formellement à une borne supérieure de confiance sur la vraie moyenne μ_a de niveau de confiance $1/T^4$, par l'inégalité de Hoeffding. D'où le nom de l'algorithme.

Concrètement, `UCB` suit à-peu-près les bras ayant la meilleure moyenne empirique $\widehat{\mu}_a(t)$, mais en accordant le bénéfice du doute aux bras qui ont été tirés peu de fois : ceux pour lesquels $N_a(t)$ est faible. On dit qu'`UCB` est un algorithme optimiste façe à l'incertitude, un principe puissant qui a inspiré de nombreux autres algorithmes de bandits.

Le principe de l'optimisme face à l'incertitude est formulé dans Lai and Robbins [1985], et des versions moins abouties de stratégies à indices sont proposées dès Agrawal [1995b] et Burnetas and Katehakis [1996]. Tous ces articles offrent des analyses exclusivement asymptotiques des stratégies proposées. C'est dans Auer et al. [2002a], qu'apparaît l'analyse moderne, c'est-à-dire non-asymptotique, de l'algorithme.

Présentons rapidement les garanties que l'on obtient avec cette stratégie. Ces garanties se séparent en deux types : les bornes qu'on qualifie de "*distribution-dependent*", et celles dites "*distribution-free*". Ces deux familles de bornes donnent naturellement lieu à deux notions d'optimalité.

Borne *distribution-dependent*

La première borne sur le regret d'`UCB` que l'on évoque généralement est une borne distribution-dependent, appelée ainsi parce qu'elle s'exprime en fonction du problème de bandit auquel le joueur fait face.

Théorème 1.1 (Auer et al. [2002a]). *Pour tout horizon de temps $T \geqslant 1$, pour n'importe quel problème de bandit $\underline{\nu}$ à paiements bornés dans $[0,1]$, si le joueur suit la stratégie `UCB`, alors pour tout bras a sous-optimal,*

$$\mathbb{E}\big[N_a(T)\big] \leqslant 8\frac{\ln T}{\Delta_a^2} + 2\,. \tag{1.2}$$

On déduit de ce résultat une borne sur le regret en appliquant la décomposition (1.1) :

$$R_T(\underline{\nu}) = \sum_{a=1}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big] \leqslant \sum_{a=1}^{K} 8\frac{\ln T}{\Delta_a} + 2\sum_{a=1}^{K} \Delta_a\,. \tag{1.3}$$

La dépendance de la borne en l'horizon de temps T est logarithmique, et on dit parfois que le regret d'`UCB` croît lentement avec le temps. Néanmoins, il faut tempérer ce propos et examiner la

dépendance en le problème. Précisément, si certains bras $a \in [K]$ sont tels que $\Delta_a \ll (\ln T)/T$, alors la majoration est plus grande que la borne triviale $R_T \leqslant T$, et elle ne nous renseigne alors pas sur le regret. Ainsi, bien qu'elle soit valable en tout temps, cette garantie est surtout intéressante pour des valeurs suffisamment grandes de T.

Borne *distribution-free*

La deuxième famille de bornes traditionnellement proposées en bandits et celle des bornes dites "distribution-free", ou minimax. Ce sont des majorations uniformes du regret sur toute la classe étudiée.

Théorème 1.2 (Audibert and Bubeck [2009]). *Pour tout horizon de temps $T \geqslant 1$, le regret d'`UCB` sur les problèmes de bandits bornés est majoré par*

$$\sup_{\underline{\nu} \in \mathcal{E}_{[0,1]}} R_T(\underline{\nu}) \leqslant 4\sqrt{2}\sqrt{KT \ln T} + 2K\,. \tag{1.4}$$

La dépendance en T est beaucoup moins bonne que celle de la borne distribution-dependent, puisqu'elle est essentiellement de l'ordre de $\sqrt{T}$ plutôt qu'en $\ln T$. La différence importante est que la borne ne dépend pas du problème en question, et en particulier, pas des écarts.

La première preuve de la borne distribution-free pour `UCB`, et pour les bandits stochastiques en général, apparaît (à ma connaissance) dans Audibert and Bubeck [2009]. La preuve illustre bien un aspect important de la théorie des bandits : la difficulté d'un problème se caractérise au niveau des bras ayant un écart faible, et l'écart critique est d'ordre $\sqrt{K/T}$.

Preuve. Le résultat découle de la borne distribution-dependent (1.2). Pour le prouver, on sépare les bras en deux groupes, selon que leur écart au meilleur bras soit grand ou non. Donnons-nous donc un seuil $\delta > 0$, dont on spécifiera la valeur plus tard. Alors,

$$R_T(\underline{\nu}) = \sum_{a=1}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big] = \sum_{\substack{a=1 \\ \Delta_a \leqslant \delta}}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big] + \sum_{\substack{a=1 \\ \Delta_a > \delta}}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big]\,. \tag{1.5}$$

On borne différemment chacune de ces deux sommes. Pour la première, qui correspond aux bras ayant un petit Δ_a, on utilise simplement la borne $\Delta_a \leqslant \delta$, puis le fait que la somme des $N_a(T)$ vaut T.

$$\sum_{\substack{a=1 \\ \Delta_a \leqslant \delta}}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big] \leqslant \sum_{\substack{a=1 \\ \Delta_a \leqslant \delta}}^{K} \delta\, \mathbb{E}\big[N_a(t)\big] \leqslant T\delta \tag{1.6}$$

Pour la seconde somme, on applique d'abord la garantie distribution-dependent, ce qui fait apparaître un facteur $1/\delta$, en utilisant le fait que $1/\Delta_a \leqslant 1/\delta$. On majore chaque terme de la somme ainsi, d'où le facteur K supplémentaire.

$$\sum_{\substack{a=1 \\ \Delta_a > \delta}}^{K} \Delta_a \mathbb{E}\big[N_a(t)\big] \leqslant \sum_{\substack{a=1 \\ \Delta_a > \delta}}^{K} \Delta_a \left(8\frac{\ln T}{\Delta_a^2} + 2\right) \leqslant \sum_{\substack{a=1 \\ \Delta_a > \delta}}^{K} 8\left(\frac{\ln T}{\Delta_a} + 2\right) \leqslant 8\frac{K \ln T}{\delta} + 2K \tag{1.7}$$

On a donc montré

$$R_T(\underline{\nu}) \leqslant T\delta + 8K\frac{\ln T}{\delta} + 2K\,. \tag{1.8}$$

Le seuil δ étant un paramètre de l'analyse, on peut l'optimiser pour obtenir une borne la plus petite possible. Le choix $\delta = \sqrt{8K \ln T/T}$ donne le résultat annoncé. □

1.1.3. Bornes inférieures et optimalité

Une fois ces bornes supérieures obtenues, le statisticien se demande naturellement s'il est possible de les améliorer : c'est la question des bornes inférieures. Nous présentons ici les bornes inférieures analogues aux garanties présentées ci-dessus, en commençant par les bornes distribution-dependent.

Borne inférieure distribution-dependent

Algorithmes uniformément convergents. Pour obtenir une borne inférieure sur le regret distribution-dependent, il faut faire une hypothèse d'uniformité sur l'algorithme que l'on considère. En effet, la stratégie absurde qui consiste à tirer uniquement le premier bras obtiendra un regret nul, et donc non-améliorable, sur tout problème pour lequel le premier bras est optimal. Pourtant, on ne saurait recommander cette stratégie, qui subit un regret catastrophique dès que le premier bras est sous-optimal. Une solution consiste à ne considérer que des stratégies dites *uniformément convergentes sur une classe d'environnements* $\mathcal{E}$. Ce sont les stratégies ψ telles que

$$\text{pour tout } \underline{\nu} \in \mathcal{E}\,, \quad \text{pour tout } \alpha > 0\,, \quad \liminf_{T\to\infty} \frac{R_T(\psi, \underline{\nu})}{T^\alpha} = 0\,. \tag{1.9}$$

On sait par exemple que l'algorithme `UCB` est uniformément convergent sur l'environnement des bandits bornés $\mathcal{E}_{[0,1]}$, d'après le Théorème 1.1.

En restreignant notre attention aux algorithmes uniformément convergents, il devient possible de quantifier la difficulté d'un problème $\underline{\nu}$ à l'intérieur de la classe $\mathcal{E}$, grâce à des outils de théorie de l'information. Grossièrement, le raisonnnement est le suivant.

Fixons un problème de bandit $\underline{\nu}$, dans une classe d'environnement $\mathcal{E}$ connue du joueur ; fixons aussi une stratégie uniformément convergente. Puisque l'algorithme est uniformément convergent, il doit être capable de distinguer le problème $\underline{\nu}$ de tous les problèmes $\underline{\nu}' \in \mathcal{E}$ n'ayant pas les mêmes actions optimales. Il doit donc accumuler de l'information utile à la différentiation entre $\underline{\nu}$ et $\underline{\nu}'$. Une grandeur qui mesure cette information est la divergence de Kullback-Leibler. On rappelle la définition de la divergence de Kullback-Leibler entre deux mesures de probabilité $\mathbb{P}$ et $\mathbb{Q}$:

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \begin{cases} \displaystyle\int \ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} & \text{si} \quad \mathbb{P} \ll \mathbb{Q}\,, \\ +\infty & \text{sinon.} \end{cases} \tag{1.10}$$

Une quantité cruciale dans les bornes inférieures est la divergence de Kullack-Leibler entre les lois des choix et des observations $(A_1, Y_1, \dots, A_T, Y_T)$ lorsque le problème est $\underline{\nu}$ ou $\underline{\nu}'$, que l'on note

$$\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^T,\, \mathbb{P}_{\underline{\nu}'}^T\big)\,. \tag{1.11}$$

Une stratégie uniformément convergente doit garantir que cette quantité d'information soit suffisamment grande pour tout $\underline{\nu}'$ n'ayant pas les mêmes bras optimaux que $\underline{\nu}$. Pour accumuler cette information, le joueur est amené à tirer des bras sous-optimaux. Précisément, une condition nécessaire pour qu'un algorithme soit uniformément convergent est que pour tout $\underline{\nu}' \in \mathcal{E}$ n'ayant pas les mêmes bras optimaux que $\underline{\nu}$ (voir par exemple le lemme 5.2 au Chapitre 5)

$$\liminf_{T\to+\infty} \frac{\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^T,\, \mathbb{P}_{\underline{\nu}'}^T\big)}{\ln T} \geqslant 1\,. \tag{1.12}$$

D'où la borne inférieure suivante sur le regret.

Une façon commode d'énoncer la borne dans le cadre d'un environnement non-structuré est d'introduire une quantité appelée la $\mathcal{K}_{\inf}$. Soit $\mathcal{D}$ un ensemble de mesures de probabilité sur $\mathbb{R}$ admettant toutes un premier moment, soient $\nu \in \mathcal{D}$ et $\mu \in \mathbb{R}$ tel que $\mathrm{E}(\nu) \leqslant \mu$,

$$\mathcal{K}_{\inf}(\nu, \mu; \mathcal{D}) \stackrel{\text{def}}{=} \inf_{\substack{\nu' \in \mathcal{D} \\ \mathrm{E}(\nu) > \mu}} \mathrm{KL}(\nu, \nu')\,, \tag{1.13}$$

On a la borne inférieure suivante sur le regret des algorithmes uniformément convergents pour les bandits bornés.

Théorème 1.3 (Lai and Robbins [1985], Burnetas and Katehakis [1996]). *Soit ψ une stratégie uniformément convergente sur la classe $\mathcal{D}^K$. Alors pour tout problème $\underline{\nu} \in \mathcal{D}^K$,*

$$\liminf_{T \to +\infty} \frac{R_T(\underline{\nu}, \psi)}{\ln T} \geqslant \sum_{\substack{a=1 \\ \Delta_a > 0}}^{K} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star;\, \mathcal{D})}\,. \tag{1.14}$$

Dans le cas des bandits bornés dans $[0,1]$, comparons cette borne inférieure à (1.3), en faisant appel à l'inégalité de Pinsker. Celle-ci garantit que si ν et ν' sont deux mesures de probabilité sur $[0,1]$, alors la divergence de Kullback-Leibler entre ces deux mesures est minorée par

$$\mathrm{KL}(\nu, \nu') \geqslant 2\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2\,. \tag{1.15}$$

En particulier, pour une action a sous-optimale dans un problème $\underline{\nu}$, si ν' est telle que $\mathrm{E}(\nu') > \mu^\star$, alors d'après l'inégalité de Pinsker,

$$\mathrm{KL}(\nu_a, \nu') > 2(\mu_a - \mu^\star)^2 = \Delta_a^2\,, \quad \text{d'où} \quad \mathcal{K}_{\inf}\big(\nu_a, \mu^\star; \mathcal{D}_{[0,1]}\big) \geqslant 2\Delta_a^2\,,$$

en prenant l'infimum sur ν'.

Ainsi, en comparant les facteurs devant le logarithme dans la borne supérieure (1.3) et la borne inférieure, on a deux termes

$$8 \sum_{\substack{a=1 \\ \Delta_a > 0}}^{K} \frac{1}{\Delta_a} \quad \text{vs.} \quad \sum_{\substack{a=1 \\ \Delta_a > 0}}^{K} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{[0,1]})}\,, \tag{1.16}$$

et l'on sait que le terme de gauche est plus grand que celui de droite. L'inégalité est d'ailleurs stricte, notamment à cause du facteur 16 qui les sépare dans la majoration par l'inégalité de Pinsker. Le statisticien souhaitera donc améliorer l'une ou l'autre de ces bornes, soit en proposant un meilleur algorithme qu'`UCB`, soit en trouvant des bornes inférieures plus fines.

C'est du côté des algorithmes que l'écart a été comblé. L'optimalité asymptotique des bornes inférieures distribution-dependent dans des modèles paramétriques était connue dès Lai and Robbins [1985], mais c'est dans Honda and Takemura [2011] qu'apparaît la première stratégie qui égale la borne inférieure pour la classe $\mathcal{E}_{[0,1]}$. Le Chapitre 2 de cette thèse est d'ailleurs consacré à une variante d'un algorithme, `KL-UCB`, qui comble aussi cet écart.

On parlera ainsi d'algorithme asymptotiquement optimal, ou distribution-dependent optimal pour une certaine classe non-structurée $\mathcal{D}^K$, lorsqu'un algorithme atteint la borne inférieure du Théorème 1.3.

Borne inférieure distribution-free

Parallèlement aux garanties distribution-dependent, discutons de l'optimalité des bornes distribution-free, ou minimax. A cette fin, énonçons la borne inférieure correspondante.

Théorème 1.4 (Auer et al. [2002b]). *Pour tout horizon de temps $T \geqslant 1$, pour tout algorithme de bandits, il existe un problème dans $\mathcal{E}_{[0,1]}$ sur lequel l'algorithme subit un regret minoré par*

$$\inf_{\psi \text{ algorithmes}} \sup_{\underline{\nu} \in \mathcal{E}_{[0,1]}} R_T(\psi, \underline{\nu}) \geqslant \frac{1}{20}\sqrt{KT}\,. \tag{1.17}$$

La preuve de ce théorème repose sur un raisonnement légèrement différent de celle des bornes inférieures distribution-dependent. Pour les bornes distribution-dependent, on fixait un algorithme et un problème de bandit, et on étudiait les conséquences de l'hypothèse de convergence sur le regret pour ce problème fixé. Dans le cas présent, l'argument consiste à identifier une famille de problèmes difficiles à l'intérieur de la classe étudiée, ici $\mathcal{E}_{[0,1]}$, et à montrer qu'aucun algorithme ne peut faire mieux que la borne énoncée sur tous ces problèmes.

Les problèmes de bandits les plus difficiles au sens minimax pour $\mathcal{E}_{[0,1]}$ sont des problèmes où il faut trouver un unique bras légèrement meilleur que les autres. Précisément, on définit une famille de K problèmes de bandits à K bras, de la façon suivante. Toutes les distributions de tous les bras seront Bernoulli (on définit donc bien des problèmes bornés), et le paiement moyen du bras a dans le i-ème problème sera :

$$\mu_a^{(i)} = \frac{1}{2} + \sqrt{\frac{K}{T}}\mathbb{1}_{\{a=i\}}\,. \tag{1.18}$$

Intuitivement, n'importe quel algorithme faisant face à de tels problèmes ne pourra pas accumuler suffisamment d'information pour distinguer le meilleur bras des autres, et sera donc obligé d'explorer (plus ou moins) uniformément. Ainsi, il subira presque à chaque tour un regret de l'ordre de $\sqrt{K/T}$, soit un regret cumulé sur les T tours de $T\sqrt{K/T} = \sqrt{KT}$.

Ce résultat se montre rigoureusement en manipulant les divergences de Kullback-Leibler entre les distributions des lois des observations.

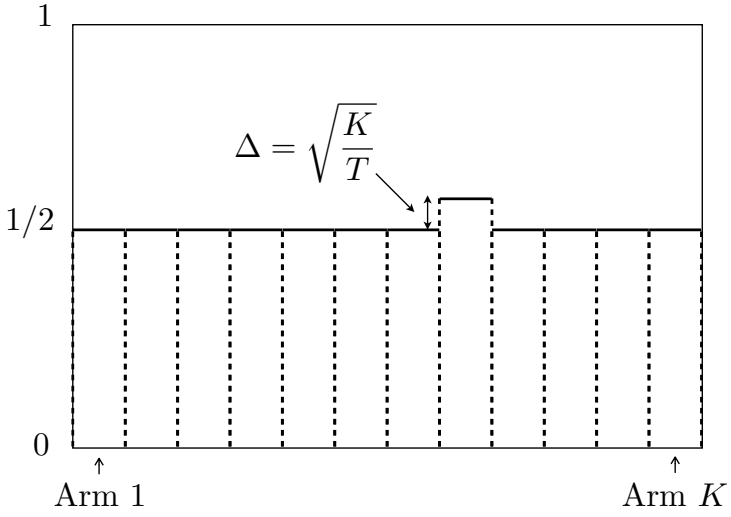


FIGURE 1.1. : Les paiements moyens des problèmes de bandits difficiles au sens minimax, cf. (1.18)

.

On dit parfois que les bornes minimax sont pessimistes, parce qu'elles ne se comparent qu'aux problèmes les plus difficiles. Ainsi, un algorithme qui obtiendrait un regret qui vaut exactement

$\sqrt{KT}$ sur tous les problèmes de bandit, serait optimal au sens minimax, mais certainement pas intéressant en pratique, puisque l'on peut obtenir un regret de l'ordre de $\ln T$ lorsque tous les écarts sont grands, par exemple avec `UCB`.

Une fois encore, il y a un écart entre la borne supérieure distribution-free d'`UCB` et la borne inférieure. La première était d'ordre $\sqrt{KT \ln T}$, tandis que la dernière était d'ordre $\sqrt{KT}$. On pourrait donc chercher soit à trouver une plus grande borne inférieure, soit à construire un algorithme bénéficiant de meilleures bornes supérieures. L'algorithme `MOSS`, de Audibert and Bubeck [2009] garantit un regret qui colle à cette borne inférieure, à un facteur numérique multiplicatif près : c'est un algorithme minimax optimal. On discute d'autres algorithmes minimax optimaux dans le Chapitre 2.

Importance du modèle

Les bornes inférieures et les notions d'optimalité dépendent fortement du modèle dans lequel on se place, c'est-à-dire de l'information que l'on suppose préalablement disponible. En général, le joueur voudra faire le moins d'hypothèses possibles. Dans la suite, on proposera plusieurs manières de mettre en œuvre ce principe. Une première manière naturelle est de s'intéresser à des environnements les plus grands possibles, puis de chercher les vitesses optimales du regret sur ces environnements.

En ce sens, l'environnement $\mathcal{E}_{[0,1]}$ est particulièrement pertinent. Au lieu de faire des hypothèses paramétriques, par exemple, supposer que les paiements suivent une loi de Bernoulli ou gaussienne, on considère le modèle non-paramétrique de toutes les distributions bornées dans $[0, 1]$.

Aussi, le Chapitre 2 de cette thèse est consacré à l'étude fine du regret optimal pour ce modèle. Dans ces conditions, coller aux bornes inférieures donne lieu à certaines difficultés techniques, dont nous discutons dans la section suivante et dans le chapitre en question.

1.1.4. Un algorithme doublement optimal pour les bandits bornés

Dans cette section, nous décrivons le regret optimal pour l'environnement des bandits à paiements bornés dans $[0, 1]$. Nous proposons en particulier un algorithme à la fois asymptotiquement optimal et minimax optimal pour ce modèle.

On se concentrera dans cette section sur l'environnement $\mathcal{E}_{[0,1]}$ des problèmes bornés dans $[0, 1]$, et $\mathcal{K}_{\inf}$ désignera toujours la $\mathcal{K}_{\inf}$ associée au modèle $\mathcal{D}_{[0,1]}$.

Affiner les bornes de confiance grâce à la $\mathcal{K}_{\inf}$

En un sens, les bornes supérieures de confiance utilisées dans l'algorithme `UCB` correspondent à des queues de distributions gaussiennes. Or, l'hypothèse de bornitude des paiements est beaucoup plus forte qu'une hypothèse de (sous-)gaussianité. Pour des distributions de paiements bornées, on s'attend donc à pouvoir affiner ces bornes de confiance, et à transformer ces bornes de confiances en un meilleur algorithme. Ce programme a été esquissé et partiellement mené à bien par Cappé et al. [2013], qui introduisent l'algorithme `KL-UCB`, défini par l'indice

$$U_a^{\mathrm{KL}}(t) = \sup\left\{u \in [0,1] :\ \mathcal{K}_{\inf}(\widehat{\nu}_a(t), u) \leqslant \frac{\ln T}{N_a(t)}\right\} \quad \text{où} \quad \widehat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{1}_{\{A_s = a\}}\,.$$

La mesure $\widehat{\nu}_a(t)$ est la mesure empiririque des observations associées au bras a. Cette formule pour les indices correspond à une borne supérieure de confiance sur la vraie moyenne μ_a. Pour le voir, commençons par la réécriture suivante, qui découle de la définition de l'indice,

$$\mathbb{P}\big[U_a^{\mathrm{KL}}(t) > \mu_a\big] \leqslant \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu_a\big) > \frac{\ln T}{N_a(t)}\right]\,. \tag{1.19}$$

Or, si $X_1, \dots, X_n$ sont n variables i.i.d. de loi ν à support dans $[0,1]$, alors en notant $\widehat{\nu}_n$ leur distribution empirique, on a l'inégalité de déviation suivante

$$\mathbb{P}\big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) > u)\big] \leqslant \mathrm{e}(2n+1)\mathrm{e}^{-nu}\,. \tag{1.20}$$

Ainsi, l'indice est, au moins formellement, une borne supérieure de confiance sur la vraie moyenne $\mathrm{E}(\nu_a)$. L'algorithme `KL-UCB` tire à chaque tour un bras maximisant cet indice.

Agrawal [1995b] introduit une première version de cet algorithme, et s'intéresse exclusivement aux garanties asymptotiques. C'est dans Cappé et al. [2013] qu'apparaît la formulation actuelle de la stratégie ; des bornes distribution-dependent optimales y sont prouvées pour certains problèmes particuliers. Dans le Chapitre 2, nous montrons que `KL-UCB` est bien asymptotiquement optimal pour le modèle entier des bandits à supports dans $[0,1]$. Nous y proposons par ailleurs une amélioration de la stratégie, qui permet d'atteindre en plus l'optimalité minimax.

Optimalité minimax et `KL-UCB-switch`

Pour améliorer les garanties distribution-free d'`UCB`, Audibert and Bubeck [2009] introduisent l'algorithme `MOSS`, qui repose sur une autre modification de l'indice `UCB` :

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2}{N_a(t)} \ln\!\left(\max\!\Big(\frac{T}{KN_a(t)}, 1\Big)\right)}\,. \tag{1.21}$$

Du fait de la modification du niveau de confiance (qui se lit à l'intérieur du logarithme), `MOSS` explore légèrement moins que `UCB`, et atteint une meilleure borne minimax.

Théorème 1.5 (Audibert and Bubeck [2009]). *`MOSS` bénéficie de la garantie distribution-free suivante :*

$$\sup_{\nu \in \mathcal{E}_{[0,1]}} R_T \leqslant 18\sqrt{KT}\,. \tag{1.22}$$

La valeur de la constante devant le $\sqrt{KT}$ est améliorée par rapport à la référence originale. Nous présentons une preuve simplifiée donnant cette constante dans le Chapitre 2, Proposition 2.2.

La modification du niveau de confiance dans `UCB` permet donc une amélioration directe de la borne distribution-free. Il est alors naturel de modifier le niveau de confiance dans l'indice `KL-UCB` pour chercher à en faire un algorithme minimax optimal : c'est l'algorithme `KL-UCB++` de Ménard and Garivier [2017],

$$U_a^{\mathrm{KL}\,++}(t) = \sup\left\{u \in [0,1] :\ \mathcal{K}_{\inf}(\widehat{\nu}_a(t), u) \leqslant \frac{1}{N_a(t)} \ln\Big(\frac{T}{KN_a(t)}\Big)\right\}\,. \tag{1.23}$$

Hélas l'inégalité de déviation (1.20) pour la KL n'est pas suffisamment forte pour obtenir les garanties espérées sur le regret de `KL-UCB++`. Nous proposons donc l'algorithme `KL-UCB-switch`, qui mélange les deux types d'indices.

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

où $f(t,K) = (t/K)^{1/5}$. L'algorithme joue donc selon `KL-UCB++`, jusqu'à ce que certains bras soient tirés suffisament de fois. Lorsqu'un bras a été tiré plus de $f(T,K)$ fois, l'algorithme attribue alors à ce bras l'indice `MOSS`. On peut justifier cette approche intuitivement en disant que l'indice `MOSS` est un indice adapté au modèle gaussien, et que les fluctuations de la distribution empirique des paiements d'un bras autour de sa moyenne se rapprochent de fluctuations gaussiennes dès lors que ce bras sera suffisamment tiré.

`KL-UCB-switch` bénéficie des garanties suivantes :

Théorème 1.6 (Théorèmes 2.1 et 2.2, Chapitre 2). *Pour tout T, l'algorithme* `KL-UCB-switch` *vérifie à la fois :*

$$\sup_{\underline{\nu}\in\mathcal{E}_{[0,1]}} R_T(\underline{\nu}) \leqslant 23\sqrt{KT} + (K-1)\,, \tag{1.24}$$

et pour tout $\underline{\nu} \in \mathcal{E}_{[0,1]}$, pour tout bras sous-optimal $a \in [K]$,

$$\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o(\ln T)\,. \tag{1.25}$$

Cet algorithme est donc optimal pour les deux types de garanties considérées. Ceci montre en particulier qu'il est bien possible d'être optimal des deux points de vue, pour l'environnement $\mathcal{E}_{[0,1]}$, ce qui n'avait rien d'évident *a priori*.

Une question alors naturelle est de savoir si l'on pourrait se passer du changement d'indice dans `KL-UCB-switch`, et ne garder que l'indice de `KL-UCB++`.

Perspectives 1.1. *On est naturellement amenés à conjecturer que l'algorithme* `KL-UCB++` *est minimax optimal, sans la modification de l'indice avec l'indice* `MOSS`*. Pour montrer ce résultat, l'obstacle technique à surmonter dans notre approche est la suppression du facteur $2n+1$ dans l'inégalité de déviation* (1.20)*.*

1.1.5. Un autre cadre de bandits préservant la diversité

Dans le Chapitre 5, un peu à part, nous étudions une modification du cadre standard des problèmes de bandits, qui permet d'incorporer des contraintes de diversité dans les choix du joueur.

Le constat initial, dressé dans Celis et al. [2019], est que les (bons) algorithmes de bandits ont tendance à se polariser, c'est-à-dire à proposer quasi-exclusivement la meilleure action. Ce comportement, souhaitable dans le cadre classique, peut être dommageable lorsque l'on a certaines applications en vue.

Par exemple, un restaurateur, suivant à la lettre un algorithme de bandits pour émettre des recommandations à chaque client, risquerait de ne proposer plus que du poisson à tous ses clients. Ceci serait désagréable pour un cuisinier aimant aussi préparer sa ratatouille ; il faudrait donc imposer au restaurateur une certaine diversité dans ses recommendations. Voir le chapitre en question pour des exemples moins farfelus, et une discussion de la littérature sur ces sujets liés à la question de la "fairness", ou "équité" des algorithmes.

Diversity-preserving bandits : cadre formel

Comme dans les bandits standards, K actions sont disponibles, l'action a générant un paiement de loi ν_a lorsque a est sélectionnée. La différence tient dans le fait que le joueur doit choisir une loi de probabilité $\underline{p}_t$ sur l'ensemble des actions $[K]$, parmi un ensemble de probabilités $\mathcal{P}$ donné, et tirer A_t selon $\underline{p}_t$.

Par exemple, on peut fixer un seuil $\ell \in]0, 1/K[$ et un ensemble de probabilités

$$\mathcal{P} = \big\{(p_1,\dots,p_K) \in \mathcal{S}_K \;:\; \text{pour tout } a\in[K],\quad p_a \geqslant \ell\big\}\,. \tag{1.26}$$

Dans cet exemple, à chaque tour, le joueur sélectionnera toujours n'importe quelle action a avec probabilité au moins ℓ. La notion de regret adéquate dans ce cadre devient

$$R_T^{\mathrm{div}}(\underline{\nu}) = T\max_{\underline{p}\in\mathcal{P}} \sum_{a=1}^{K} p_a\mu_a - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]\,, \tag{1.27}$$

c'est-à-dire la différence entre le paiement moyen obtenu, et celui que le joueur aurait gagné s'il n'avait joué que la meilleure probabilité.

Ce cadre, un peu différent du cadre standard, donne lieu à des phénomènes inhabituels. On a vu par exemple que dans les problèmes classiques de bandits, le regret croît toujours logarithmiquement lorsque $T \to \infty$, cf. la borne inférieure du Théorème 1.3. Dans ce nouveau cadre, nous montrons que cette vitesse logarithmique est toujours nécessaire sur certains couples $\underline{\nu}, \mathcal{P}$, mais que pour d'autres, il est possible d'atteindre un regret constant.

Perspectives 1.2. *Ce chapitre étant un travail en cours, il reste des questions auxquelles nous souhaiterions répondre. Le géométrie de l'ensemble de probabilités $\mathcal{P}$ a une influence forte sur les performances possibles des algorithmes ; l'exemple le plus frappant étant la possibilité du regret constant lorsque $\mathcal{P}$ est inclus dans l'intérieur du simplexe des probabilités sur $[K]$. Ainsi, nous aimerions étudier plus finement la dépendance en $\mathcal{P}$, et déceler les caractéristiques de $\mathcal{P}$ qui détermine le regret optimal. Précisément, on pourrait par exemple chercher à calculer le regret minimax en fonction de $\mathcal{P}$. Plusieurs conjectures précises qui portent sur ce regret optimal sont formulées au Chapitre 5.*

1.2. Introduction aux problématiques d'adaptation par un exemple

Dans cette section, on présente un exemple de problème de bandits à ensemble d'actions continu. Cet exemple nous servira de fil conducteur pour introduire nos résultats des Chapitres 3 et 4.

1.2.1. Chercher de l'or dans une rivière : un problème de bandits continus

Les exemples de problèmes de bandits décrits précédemment sont des cas particuliers de problèmes dit discrets (ou à ensemble de bras finis). L'image typiquement utilisée est que le statisticien est un joueur de casino, et que chaque bras est une machine à sous. Imaginons cette fois que le joueur, écrasé par ses dettes dues au casino, décide de se reprendre en main et de partir chercher de l'or dans une rivière de l'Oklahoma.

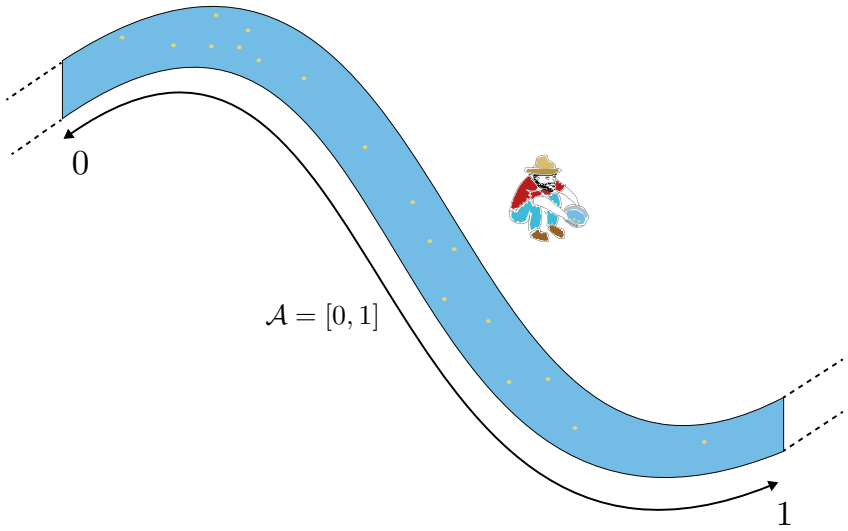


FIGURE 1.2. : Une rivière aurifère et un chercheur (d'or).

Le statisticien-orpailleur procède de la manière suivante. Il fixe un grand tronçon de rivière et indexe chaque point de la rivière par sa coordonnée naturelle $a \in [0, 1]$. Chaque jour, il choisit

un site $A_t \in [0,1]$, sur lequel il passera sa journée à chercher de l'or. Il récolte alors une certaine quantité d'or Y_t, qui sera un réel positif $Y_t \in \mathbb{R}_+$ (mesurée par exemple en milligrammes d'or).

Si l'on fait l'hypothèse de modélisation que le paiement Y_t, conditionnellement au lieu A_t suit une loi de probabilité fixée ν_{A_t}, et est indépendant des paiements passés (hypothèse bien entendu discutable, mais que nous ferons pour l'exemple), alors le statisticien joue à un jeu de bandits continus. Le nom bandits continus fait référence au fait que l'espace des actions, ici la rivière, ou l'intervalle $[0,1]$ est un ensemble infini indénombrable. Pour éviter la confusion avec l'ensemble de valeurs des paiements, on note l'ensemble des actions $\mathcal{A} = [0,1]$.

On suppose donc qu'il existe une famille de mesures de probabilité $\underline{\nu} = (\nu_a)_{a\in\mathcal{A}}$, qui modélise la distribution de l'or récolté par le statisticien au cours de ses recherches. Notons aussi la fonction de paiements-moyens associée au problème $\underline{\nu}$,

$$f_{\underline{\nu}} : a \mapsto \mathrm{E}(\nu_a)$$

(on supposera toujours que toutes les distributions de paiements ont un premier moment fini).

Le statisticien cherche une stratégie lui permettait de récolter le plus d'or possible. On définit le regret d'une stratégie ψ sur le problème $\underline{\nu}$ de la façon suivante

$$R_T(\psi, \underline{\nu}) = T \max_{a\in\mathcal{A}} f_{\underline{\nu}}(a) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] .$$

Quelle stratégie le statisticien doit-il adopter pour récolter le plus d'or possible ? La réponse dépendra des hypothèses que l'on fait sur le problème $\underline{\nu}$, c'est-à-dire des connaissances préalables dont dispose le joueur sur la répartition de l'or dans la rivière. Présentons dans un premier temps un certain jeu d'hypothèses, que nous relacherons par la suite.

1.2.2. Un jeu d'hypothèses : support des paiements et régularité des paiements moyens fixés

Le joueur commence par faire des hypothèses sur le problème $\underline{\nu}$. D'une part, il sait grâce à l'expérience de collègues orpailleurs qu'il n'obtiendra jamais plus d'un milligramme d'or en une journée sur cete rivière. Formellement, on impose donc que toutes les distributions des paiements soient à support dans $[0,1]$.

D'autres collègues lui ont aussi fait part de l'observation suivante : la différence de paiement moyen (en mg) entre deux points n'excède jamais la distance (rapportée à $[0,1]$) entre ces points. Autrement dit, la fonction de paiements-moyens est (1–)Lipschitzienne :

$$\text{pour tous } x, y \in \mathcal{A}\,, \quad \left|f_{\underline{\nu}}(x) - f_{\underline{\nu}}(y)\right| \leqslant |x - y| \ .$$

Il s'agit d'une hypothèse de régularité sur la fonction de paiements-moyens. Une autre manière d'interpréter cette hypothèse est que le joueur, en sélectionnant l'action a, récupère aussi de l'information sur les paiements des actions proches de a.

On résume ces hypothèses en disant que le joueur fait face à un problème de bandits qui appartient à la classe

$$\mathcal{E}^{\mathrm{Lip}} = \left\{(\nu_a)_{a\in\mathcal{A}} : \text{ pour tout } a \in \mathcal{A}\,, \quad \nu_a\big([0,1]\big) = 1, \quad \text{et} \quad f_{\underline{\nu}} \text{ est Lipschitzienne}\right\} .$$

Décrivons maintenant une stratégie simple mais efficace pour ce problème continu : la discrétisation.

Discrétisations

Les vitesses minimax sont atteintes en découpant l'ensemble des actions $\mathcal{A}$ en K morceaux de même taille, et en choisissant correctement K. L'idée, proposée dans Kleinberg [2004] avec l'algorithme `CAB1` (pour Continuum-Armed Bandits), consiste donc à discrétiser l'ensemble des actions, et à jouer dans les morceaux discrétisés selon les recommendations d'un algorithme discret. Présentons plus précisément l'approche.

Fixons un problème de bandits $\nu \in \mathcal{E}^{\mathrm{Lip}}$, et notons f sa fonction de paiements-moyens. Soit $K \geqslant 2$ un entier, qui désignera le nombre de cellules de discrétisations, et soit ψ^{disc} un algorithme de bandits à K bras, qu'on appellera l'algorithme discret.

Divisons l'ensemble des actions $\mathcal{A} = [0,1]$ en K morceaux de même taille, $[(i-1)/K,\, i/K[$ (il n'est pas nécessaire que les intervalles forment une partition de $\mathcal{A}$). A chaque tour, l'algorithme discret ψ^{disc} recommande une action $I_t \in [K]$, et le joueur tire un bras uniformément au hasard dans le I_t-ème intervalle de la discrétisation. Il observe alors un paiement Y_t, qu'il transmet à l'algorithme discret.

Si le joueur choisit $I_t \in [K]$ et tire un bras A_t distribué uniformément dans $\big[(I_t-1)/K,\, I_t/K\big]$, le paiement moyen reçu vérifie alors

$$\mathbb{E}\big[Y_t \mid I_t\big] = \mathbb{E}\big[f(A_t) \mid I_t\big] = \int_{(I_t-1)/K}^{I_t/K} f(x)\,\mathrm{d}x \overset{\mathrm{def}}{=} m_{I_t}(f)\,.$$

Ainsi, du point de vue de l'algorithme discret, le paiement moyen associé au i-ème intervalle est la valeur moyenne de f sur cet intervalle, que l'on note $m_i(f)$.

Faisons une analyse rapide de cette stratégie. Soit $x^\star$ un point où f atteint son maximum. Le regret se décompose de la façon suivante :

$$R_T = T f(x^\star) - \mathbb{E}\big[f(X_t)\big] = T\bigg(f(x^\star) - \max_{i\in[K]} m_i(f)\bigg) + T \max_{i\in[K]} m_i(f) - \mathbb{E}\left[\sum_{t=1}^{T} m_{I_t}(f)\right].$$

Le deuxième terme est exactement le regret subit par l'algorithme ψ^{disc} face à un problème de bandits bornés à K bras, de moyennes $\big(m_1(f), \dots, m_K(f)\big)$. En prenant par exemple l'algorithme `MOSS` comme algorithme discret, grâce au Théorème 1.5, on obtient la garantie

$$R_T \leqslant T\bigg(f(x^\star) - \max_{i\in[K]} m_i(f)\bigg) + 18\sqrt{KT}\,.$$

Le regret est ainsi décomposé en deux termes : le premier est un terme d'approximation du maximum de f par la famille $\big(m_i(f)\big)_{i\in[K]}$, tandis que le second vient du coût de l'apprentissage du problème approximé.

En faisant un tel choix, et en considérant l'intervalle $i^\star$ contenant $x^\star$, on obtient la borne sur l'erreur d'approximation :

$$f(x^\star) - \max_{i\in[K]} m_i(f) \leqslant f(x^\star) - m_{i^\star}(f) \leqslant \frac{1}{K}\,, \quad \text{d'où} \quad R_T \leqslant \frac{T}{K} + 18\sqrt{KT}\,.$$

On peut alors choisir K de manière à minimiser cette borne supérieure : prenons par exemple $K = \lfloor T^{1/3} \rfloor$ pour obtenir

$$R_T \leqslant 2\,T^{2/3} + 18\,T^{2/3} = 20\,T^{2/3}\,.$$

Cette vitesse de croissance du regret, en $T^{2/3}$, est la vitesse minimax (optimale) sur la classe $\mathcal{E}^{\mathrm{Lip}}$ (cf. le Théorème 1.9 discuté ci-dessous). On obtient donc les meilleures vitesses minimax possible en discrétisant l'ensemble des actions.

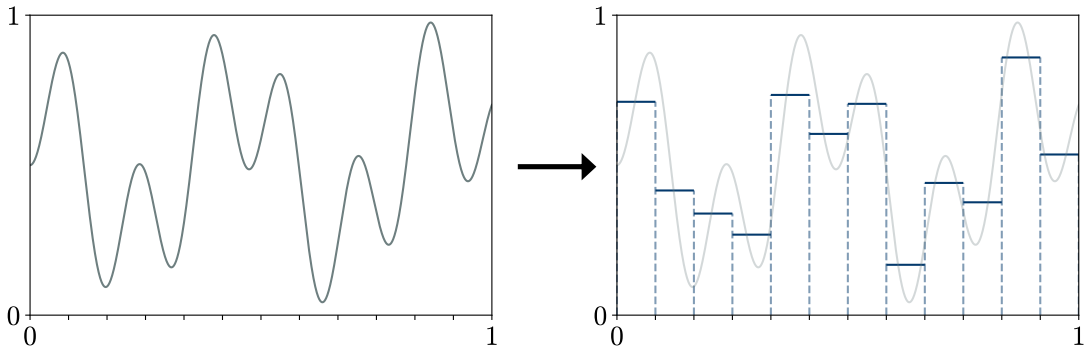


FIGURE 1.3. : Discrétisation d'un problème de bandits continus. La figure de gauche représente la fonction de paiements-moyens. En appliquant la stratégie de discrétisation, le joueur se retrouve face au problème discret représenté par les barres horizontales dans la figure de droite.

Rappelons cependant que nous avions fait deux hypothèses fortes sur les paiements : le fait que les paiements soient bornés dans l'intervalle $[0,1]$ et la connaissance précise de la régularité de la fonction de paiements-moyens. Dans la suite de cette introduction, nous discuterons de façons de nous passer de ces hypothèses.

Ces questions sont cruciales pour relier la théorie à la pratique. Très souvent, des algorithmes sont construits en supposant connus certains paramètres du problème $\underline{\nu}$. Le practicien devra alors deviner la valeur du paramètre, souvent à l'instinct : les garanties théoriques ne sont alors pas forcément valables. Les algorithmes bénéficiant de garanties adaptativess sont en ce sens plus robustes aux éventuelles incertitudes sur les hypothèses.

Commençons par nous attaquer à la première condition : la bornitude des paiements.

1.3. Adaptation minimax au support des bandits bornés

Dans ce qui précède, nous avions toujours supposé que les paiements étaient bornés dans $[0,1]$, et que le joueur disposait de cette information à l'avance. Cette hypothèse peut être irréaliste en pratique. En fait, la question se pose déjà dans les problèmes de bandits discrets ; discutons donc du problème de l'adaptation au support des paiements dans les bandits à K bras.

1.3.1. Bandits bornés à support quelconque connu

Plaçons-nous dans le cadre des bandits standards à K bras. Au lieu d'imposer que les paiements soient tous dans $[0,1]$, considérons qu'ils appartiennent tous à un intervalle $[m,M]$, où m et M sont deux réels quelconques tels que $m < M$. Supposons dans un premier temps que m et M sont connus du joueur. La classe d'environnements que l'on étudie est alors

$$\mathcal{E}_{[m,M]} = \big\{(\nu_1,\dots,\nu_K) \;:\; \forall a \in [K]\,, \quad \nu_a\big([m,M]\big) = 1\big\}\,.$$

Puisque le joueur connaît m et M, il lui suffit bien sûr de normaliser les paiements pour se ramener au cas où ils appartiennent à $[0,1]$, via la transformation

$$Y_t \leftarrow \frac{Y_t - m}{M-m} + m\,.$$

Le jeu sur les paiements transformés est complètement équivalent au jeu standard ; seul le regret est modifié, par un facteur multiplicatif $(M-m)$. En combinant les résultats discutés précédemment

pour $[0,1]$ (les Théorèmes 1.4 et 1.5), on obtient alors

$$\frac{1}{20}(M-m)\sqrt{KT} \leqslant \inf_{\psi} \sup_{\underline{\nu}\in\mathcal{E}_{[m,M]}} R_T(\underline{\nu},\psi) \leqslant 18(M-m)\sqrt{KT}\,. \tag{1.28}$$

On dira que $(M-m)\sqrt{KT}$ est la vitesse minimax optimale pour la classe $\mathcal{E}_{[m,M]}$.

Par contre, lorsque les bornes des supports des paiements sont inconnues, il n'y a plus d'opération standard permettant de se ramener à un cas connu. On pourrait d'ailleurs, au premier abord, penser que l'adaptation au support est une tâche impossible. Réfléchissons-y mieux.

1.3.2. Bandits bornés à support inconnu : adaptation (quasi-)parfaite...

Cette fois-ci, on ne suppose plus que les bornes des supports sont connues. Cela revient donc à considérer un problème qui appartient à n'importe quelle classe $\mathcal{E}_{[m,M]}$, i.e.,

$$\underline{\nu} \in \bigcup_{\substack{m,M\in\mathbb{R}\\ m<M}} \mathcal{E}_{[m,M]}$$

Notons d'ailleurs la subtile différence avec le cas où l'on ne suppose pas les bandits bornés, qui correspondrait à prendre la classe de toutes les distributions sans exception ; ce second cas est trop général pour qu'on puisse en dire des choses intéressantes.

La question que l'on se pose alors est : peut-on s'adapter aux supports des paiements ? En d'autres termes, peut-on construire un algorithme ψ garantissant pour tous $m, M \in \mathbb{R}$ tels que $m < M$,

$$\sup_{\underline{\nu}\in\mathcal{E}_{[m,M]}} R_T(\underline{\nu},\psi) \leqslant c(M-m)\sqrt{KT}\ ?$$

Reformulons encore une fois la question : on se demande si l'on pourra obtenir un algorithme garantissant pour une certaine constante numérique $c > 0$

$$\sup_{\substack{m,M\in\mathbb{R}\\ m<M}} \sup_{\underline{\nu}\in\mathcal{E}_{[m,M]}} \frac{R_T(\underline{\nu},\psi)}{(M-m)\sqrt{KT}} \leqslant c\ ?$$

Et on dira qu'un tel algorithme atteint la vitesse adaptative $(M-m)\sqrt{KT}$.

L'une des contributions de cette thèse, détaillée dans le Chapitre 4, est l'apport d'une réponse (positive) à cette question, par la construction d'algorithmes permettant effectivement d'atteindre l'adaptation aux vitesses (quasi-)minimax.

Théorème 1.7 (Théorème 4.3, Chapitre 4). *Il existe un algorithme garantissant pour tout T, pour tous m, M,*

$$\sup_{\underline{\nu}\in\mathcal{E}_{[m,M]}} R_T(\underline{\nu},\psi) \leqslant 7(M-m)\sqrt{KT\ln K} + 10(M-m)K\ln K\,. \tag{1.29}$$

L'algorithme en question repose sur des techniques utilisées dans les bandits adversariaux, un cadre cousin des bandits stochastiques dans lequel les paiements ne sont plus nécessairement identiquement distribués. On atteint ainsi les vitesses minimax classiques à un facteur multiplicatif $\sqrt{\ln K}$ près, sans la connaissance préalable du support.

1.3.3. ... mais un coût à l'adaptation dans les vitesses asymptotiques

Vu les résultats discutés dans la section précédente, on pourrait croire que l'on peut s'adapter au support des paiements sans frais. Ce n'est pas du tout le cas. Pour le voir, il faut regarder du côté des garanties distribution-dependent.

Lorsque le support est donné au joueur, on sait que celui-ci peut bénéficier d'un regret logarithmique lorsque $T \to +\infty$. Il lui suffit par exemple de jouer selon `UCB` pour les paiements normalisés. Nous avons même montré qu'il était possible d'avoir un algorithme simultanément optimal des points de vue distribution-dependent et minimax, en utilisant l'algorithme `KL-UCB-switch`.

On pourrait donc espérer obtenir aussi des bornes logarithmiques lorsque le support est inconnu. Cette poursuite est vaine : aucun algorithme n'admet de telles garanties. Nous montrons dans le Chapitre 4 une borne inférieure prouvant cette affirmation. On dit qu'un algorithme ψ atteint la vitesse adaptative $B(m, M, T)$ s'il existe une constante numérique $c > 0$ telle que pour tout horizon de temps T,

$$\sup_{\substack{m,M\in\mathbb{R}\\ m<M}} \sup_{\underline{\nu}\in\mathcal{E}_{[m,M]}} \frac{R_T(\underline{\nu},\psi)}{B(m,M,T)} \leqslant c\,. \tag{1.30}$$

On dira aussi que la vitesse est homogène si la dépendance en le support est la dépendance naturelle, i.e., s'il existe une fonction $\tilde{B}(T)$ telle que $B(m, M, T) = (M - m)\tilde{B}(T)$. La borne inférieure est la suivante.

Théorème 1.8 (Théorème 4.2, Chapitre 4). *Si un algorithme ψ atteint la vitesse adaptative homogène $(M - m)\tilde{B}(T)$ alors pour tous $m < M$ et pour tout problème $\underline{\nu} \in \mathcal{E}_{[m,M]}$, on a*

$$\liminf_{T\to+\infty} \frac{R_T(\underline{\nu},\psi)}{(M-m)\,T/\tilde{B}(T)} \geqslant \frac{1}{4}\sum_{a=1}^{K}\Delta_a\,. \tag{1.31}$$

Bien que l'adaptation au sens minimax soit possible, elle restreint les vitesses distribution-dependent accessibles ! A l'extrême, un algorithme s'adaptant au support avec un regret de l'ordre de $(M - m)\sqrt{KT}$ subira un regret sur un problème précis $\underline{\nu}$ d'au moins $\kappa(\underline{\nu})\sqrt{T}$, ce qui est bien loin d'un $\ln T$ rêvé.

1.4. Adaptation minimax à la régularité de la fonction de paiements-moyens

Revenons au problème initial de notre chercheur d'or, et essayons maintenant de nous passer de l'hypothèse de régularité de la fonction de paiements-moyens, ou plutôt de la réduire. Il est raisonnable de penser que la quantité d'or dans la rivière admet une certaine forme de régularité, et que s'il y a beaucoup d'or à un point donné, il y en aura beaucoup dans les points proches. Par contre, nous avions fixé de façon complètement arbitraire une évaluation quantitative de cette régularité, en supposant que la fonction de paiements-moyens était 1−Lipschitzienne ; c'est ce choix arbitraire que nous remettrons en cause ici.

Notons qu'il est indispensable de faire une sorte d'hypothèse de régularité sur le problème. Autrement, face à une fonction trop irrégulière, on ne pourrait rien dire ; aucun algorithme ne parviendrait à déceler un pic de paiement trop pointu.

Définissons ici une classe de régularité dérivée de la régularité Hölder. On note $\mathcal{H}(\alpha)$ l'ensemble des fonctions $f : [0, 1] \to \mathbb{R}$ qui atteignent leur maximum en un point $x^\star \in \mathcal{A}$ et telles que

$$\forall x \in \mathcal{A}\,, \quad |f(x^\star) - f(x)| \leqslant |x^\star - x|^\alpha\,.$$

Notre hypothèse de régularité est moins forte que la vraie régularité Hölder, puisqu'on ne compare les valeurs de f qu'autour de $x^\star$. Ceci nous permet d'ailleurs de considérer des valeurs $\alpha > 1$ sans nous restreindre aux fonctions constantes. On commettra parfois un léger abus de langage en parlant de régularité Hölder pour notre hypothèse. On s'intéresse alors aux problèmes de bandits à paiements bornés tels que la fonction de paiements-moyens appartient à $\mathcal{H}(\alpha)$:

$$\mathcal{E}^{\mathrm{H}}(\alpha) = \big\{(\nu_a)_{a\in\mathcal{A}} : \text{ pour tout } a \in \mathcal{A}\,, \quad \nu_a\big([0,1]\big) = 1\,, \quad \text{et} \quad f_{\underline{\nu}} \in \mathcal{H}(\alpha)\big\}\,.$$

1.4.1. La discrétisation atteint les vitesses minimax à régularité connue

Si le joueur sait à l'avance que le problème $\underline{\nu}$ contre lequel il joue appartient à $\mathcal{E}^{\mathrm{H}}(\alpha)$, avec α connu, alors il pourra appliquer la méthode de discrétisation discutée dans la section précédente, et ainsi obtenir des bornes supérieures sur le regret de l'ordre de $T^{(\alpha+1)/(2\alpha+1)}$. Il lui suffit de choisir correctement le nombre de cellules de discrétisations en fonction de α. Ces vitesses sont par ailleurs les vitesses minimax optimales sur les classes de régularité $\mathcal{E}^{\mathrm{H}}(\alpha)$.

Théorème 1.9 (Kleinberg [2004]). *Il existe des constantes numériques c et c' telles que*

$$c\,T^{(\alpha+1)/(2\alpha+1)} \leqslant \inf_{\psi} \sup_{\underline{\nu}\in\mathcal{E}^{\mathrm{H}}(\alpha)} R_T(\underline{\nu},\psi) \leqslant c'\,T^{(\alpha+1)/(2\alpha+1)}\,. \tag{1.32}$$

La borne supérieure est atteinte grâce à l'algorithme de discrétisation, avec un nombre de cellules de discrétisations bien choisi en fonction de α. Ce résultat est donné dans Kleinberg [2004], et rappelé en détail dans le Chapitre 3 (Proposition 3.2).

Que faire si l'on n'a pas d'information préalable sur α ? Encore une fois, on aimerait concevoir un algorithme obtenant les mêmes garanties minimax, mais sans la connaissance de α. La question est donc : peut-on obtenir un algorithme tel que, pour une constante numérique $c > 0$,

$$\sup_{\alpha>0} \sup_{\underline{\nu}\in\mathcal{E}^{\mathrm{H}}(\alpha)} \frac{R_T(\underline{\nu},\psi)}{T^{(\alpha+1)/(2\alpha+1)}} \leqslant c\ ? \tag{1.33}$$

(Voir l'introduction du Chapitre 3 pour une revue de littérature sur les questions d'adaptation en bandits continues.)

1.4.2. Impossibilité de l'adaptation à la régularité Hölder

Locatelli and Carpentier [2018] montrent que l'adaptation minimax aux vitesses usuelles est impossible, grâce à la borne inférieure suivante.

Théorème 1.10 (Locatelli and Carpentier [2018]). *Soit T un entier positif. Soit $B_T > 0$ un réel positif. Soient $\alpha, \gamma > 0$ deux exposants de Hölder tels que $\alpha < \gamma$.*

Supposons que $2^{-3}\,12^{\alpha} B^{-1} \leqslant T^{\alpha/2}\,2^{(1+\alpha)(8-2\gamma)}$. Si un algorithme ψ bénéficie de la garantie suivante sur la classe de régularité γ-Hölder,

$$\sup_{\underline{\nu}\in\mathcal{E}^{H}(\gamma)} R_T(\psi,\underline{\nu}) \leqslant B_T\,, \tag{1.34}$$

alors le regret minimax sur la classe α-Hölder de cet algorithme est minoré par

$$\sup_{\underline{\nu}\in\mathcal{E}^{\mathrm{H}}(\alpha)} R_T(\psi,\underline{\nu}) \geqslant 2^{-10}\,T B_T^{-\alpha/(\alpha+1)}\,. \tag{1.35}$$

Par conséquent, si un algorithme est minimax optimal sur $\mathcal{E}^{\mathrm{H}}(\alpha)$, on peut appliquer le théorème avec $B_T = cT^{(\alpha+1)/(2\alpha+1)}$, et obtenir la borne inférieure sur pour la classe de régularité $\gamma\dagger$

$$\sup_{\underline{\nu}\in\mathcal{E}^{\mathrm{H}}(\gamma)} R_T(\underline{\nu}) \geqslant c'\, T^{1-(\alpha+1)/(2\alpha+1)\alpha/(\alpha+1)} = c'\, T^{1-\alpha/(2\alpha+1)} \gg T^{(\gamma+1)/(2\gamma+1)}\,, \tag{1.36}$$

d'où l'impossibilité de garantir un regret d'ordre $T^{(\gamma+1)/(2\gamma+1)}$ sur la classe $\mathcal{E}^{\mathrm{H}}(\gamma)$.

Ceci étant établi, le joueur peut tout de même se demander quelles vitesses il est possible d'atteindre sans la connaissance du paramètre α. Afin de mener cette discussion, concentrons-nous sur l'exposant sur T dans les vitesses minimax.

Supposons que l'on dispose d'un algorithme garantissant (sans la connaissance de α) pour tout α

$$\sup_{\underline{\nu}\in\mathcal{E}^{\mathrm{H}}(\alpha)} R_T(\underline{\nu}) \leqslant cT^{\theta(\alpha)}\,, \tag{1.37}$$

où θ est une fonction (décroissante) de α, et c une constante numérique. On dira alors que l'algorithme atteint la vitesse θ. Moralement, la borne inférieure du Théorème 1.10, stipule que θ doit vérifier l'inéquation

$$\text{pour tous } \alpha,\gamma \text{ tels que } \alpha \leqslant \gamma\,, \qquad \theta(\gamma) \geqslant 1 - \frac{\alpha}{\alpha+1}\theta(\alpha)\,. \tag{1.38}$$

Tout algorithme se passant de la connnaissance du paramètre α atteindra donc une vitesse θ vérifiant l'inéquation précédente. Pour minorer cette vitesse, on peut donc chercher les solutions minimales, au sens de l'ordre ponctuel, à cette inéquation. C'est le contenu du théorème suivant.

Théorème 1.11 (Théorème 3.2, Chapitre 3). *Si un algorithme atteint la vitesse θ, alors il existe une vitesse θ_m parmi les vitesses*

$$\theta_m : \alpha \mapsto \max\left(m\,,\ 1 - m\frac{\alpha}{\alpha+1}\right)\,, \quad m \in [1/2, 1]\,. \tag{1.39}$$

telle que pour tout α, on ait $\theta(\alpha) \geqslant \theta_m(\alpha)$.

Empruntons ici une terminologie statistique standard et disons que θ_m est une famille de vitesses admissibles, c'est-à-dire des vitesses qu'il est impossible d'améliorer uniformément pour tout α. Il s'agit donc d'une famille de vitesses toutes optimales et incomparables entre elles. Ces vitesses sont représentées dans la figure 1.4.

Un algorithme admissible

Pour justifier complètement la dénomination des vitesses admissibles, encore faut-il prouver qu'elles sont bien atteintes par un algorithme. Une des contributions principales de cette thèse est justement la construction d'un algorithme capable d'atteindre n'importe laquelle des vitesses admissibles θ_m (à un facteur logarithmique près).

Théorème 1.12 (Théorème 3.2, Chapitre 3). *Pour n'importe quel $m \in [1/2, 1]$, et pour tout T, l'algorithme* `MeDZO`*, proposé au Chapitre 3, réglé avec le m choisi, atteint la vitesse admissible θ_m. Précisément, pour tout $\alpha > 0$,*

$$\sup_{\underline{\nu}\in\mathcal{E}^{H}(\alpha)} R_T(\underline{\nu}) \leqslant 412\,(\ln_2 T^m)^{3/2} T^{\theta_m(\alpha)}\,. \tag{1.40}$$

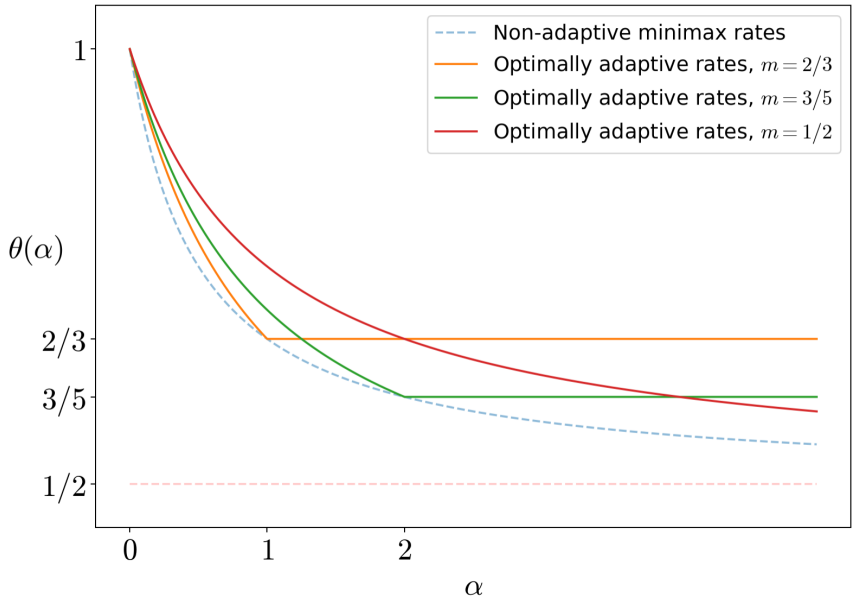


FIGURE 1.4. : La famille $(\theta_m)_{m\in[1/2,1]}$ des vitesses admissibles.

L'algorithme `MeDZO` (pour "Memorize, Discretize and Zoom Out") repose sur une idée naturelle : combiner différents niveaux de discrétisation. En revanche, l'approche concrète se distingue fortement des techniques antérieures en bandits continus. En effet, les principaux algorithmes du domaine (cf. `HOO` de Bubeck et al. [2011b] et `Zooming` de Kleinberg et al. [2019]) discrétisent à des niveaux de plus en plus fins, en affinant la discrétisation grâce à la connaissance d'un paramètre de régularité. Au contraire, `MeDZO` commence par une discrétisation fine de l'espace des bras, et dézoome au fur et à mesure que le temps passe !

Perspectives 1.3. *L'absence de bornes distribution-dependent pour les bandits continus constitue un trou dans la littérature. En général, pour des classes de problèmes suffisamment grandes, on ne peut plus espérer de bornes en* $\ln T$*, mais plutôt de l'ordre de* $\sqrt{T}$ *même en distribution-dependent, ce qui contraste avec le cas discret. Si certaines bornes de ce type sont discutées dans Kleinberg et al. [2019] dans des cas particuliers, les résultats sont loin d'égaler en finesse leur analogue du cas discret.*

Les techniques standards, qui donnent les bornes distribution-dependent dans le cas discret (cf. Théorème 1.3) ne fonctionnent plus quand l'espace des actions est trop grand.

Une étape pour obtenir des résultats plus proches du distribution-dependent serait de continuer à faire de l'adaptation minimax, mais sur des classes plus petites, en considérant d'autres notions de régularité. Par exemple, Auer et al. [2007] introduisent la régularité de marge d'un problème $\underline{\nu}$*, et Locatelli and Carpentier [2018] proposent un algorithme s'adaptant à cette régularité de marge, à condition que l'on connaisse par ailleurs la régularité Hölder. On pourrait donc chercher à s'adapter aux deux notions de régularité simultanément.*

Perspectives 1.4. *Nous avons donc étudié séparément deux problèmes : d'un côté l'adaptation aux supports des paiements pour des bandits discrets, et de l'autre l'adaptation à la régularité pour les bandits continus. Quid de l'adaptation simultanée aux supports et à la régularité dans notre problème initial ?*

Obtenir des bornes supérieures ne devrait pas être trop difficile, en combinant les stratégies proposées dans cette thèse. Il serait aussi intéressant de considérer des bornes inférieures, et de déterminer si les résultats qui découlent naturellement de notre approche suffisent à garantir des bornes optimales.

1.5. Perspectives : généralités sur l'adaptation minimax dans les problèmes de bandits

Les similarités formelles entre les questions étudiées dans les Sections 1.3 et 1.4 suggèrent une formulation générale de la problématique de l'adaptation minimax, que nous détaillons ici. En prenant un point de vue plus abstrait par rapport à la section précédente, nous énoncerons de potentiels futurs axes de recherche.

1.5.1. Adaptation minimax : formulation générale

Problèmes de bandits généraux, classes d'environnements inconnues. Généralisons d'abord la définition d'un problème de bandits, pour prendre en compte un ensemble d'actions quelconque. Soit $\mathcal{A}$ un tel ensemble. Un problème de bandit à actions dans $\mathcal{A}$ est une famille indexée par $\mathcal{A}$ de distributions de probabilité, $\underline{\nu} = (\nu_a)_{a\in\mathcal{A}}$ telles que chaque ν_a possède un premier moment fini. Une classe de problèmes $\mathcal{E}$ est tout simplement un ensemble de problèmes de bandits.

Soit $\big(\mathcal{E}(\kappa)\big)_{\kappa\in\Theta}$ une famille d'environnements, indexée par un paramètre κ vivant dans un ensemble de paramètres Θ. Le paramètre κ désignera typiquement une quantité inconnue du joueur. La classe d'environnement global auquel le joueur fait face est alors l'union des classes $\mathcal{E}(\kappa)$. Nous présenterons des exemples dans la suite.

Vitesses minimax adaptatives. On supposera typiquement que l'on connaît les vitesses minimax du regret à κ connu, c'est-à-dire pour un environnement $\mathcal{E}(\kappa)$ fixé. Ainsi, on dit que $B^\star(\kappa, T)$ est une vitesse minimax optimale sur $\mathcal{E}(\kappa)$ s'il existe des constantes numériques c_1 et c_2 telles que $0 < c_1 \leqslant c_2$ et pour tout T,

$$c_1 \leqslant \inf_{\psi} \sup_{\underline{\nu}\in\mathcal{E}(\kappa)} \frac{R_T(\underline{\nu}, \psi)}{B^\star(\kappa, T)} \leqslant c_2 \ . \tag{1.41}$$

L'inégalité de droite suppose donc qu'il existe un algorithme dont le regret est essentiellement borné par $B^\star(\kappa, T)$, tandis que celle de gauche nous dit qu'il n'existe pas d'algorithme faisant mieux. (Notons que la vitesse optimale est définie à une constante multiplicative près.)

Lorsque l'on souhaite faire de l'adaptation, on suppose que κ est inconnu, et on s'intéresse alors aux vitesses atteintes par un algorithme fixé, sur toutes les classes $\mathcal{E}(\kappa)$ simultanément. Un algorithme ψ atteint la vitesse adaptative $B^{\mathrm{ADA}}(\kappa, T)$ s'il existe une constante numérique $c > 0$ telle que

$$\sup_{\kappa\in\Theta} \sup_{\underline{\nu}\in\mathcal{E}(\kappa)} \frac{R_T(\underline{\nu}, \psi)}{B^{\mathrm{ADA}}(\kappa, T)} \leqslant c\,. \tag{1.42}$$

Evidemment, il est plus difficile de travailler à κ inconnu qu'à κ connu ; ainsi, si un algorithme atteint la vitesse adaptative $B^{\mathrm{ADA}}(\kappa, T)$, alors il existe une vitesse minimax $B^\star(\kappa, T)$ telle que pour tous T et κ,

$$B^\star(\kappa, T) \leqslant B^{\mathrm{ADA}}(\kappa, T)\,. \tag{1.43}$$

L'objectif du statisticien est d'obtenir des vitesses adaptatives les plus petites possibles. Idéalement, on espère s'adapter à la vitesse minimax $B^\star(\kappa, T)$, c'est-à-dire trouver un algorithme qui garantirait la même vitesse du regret que s'il connaissait le paramètre κ à l'avance.

En statistique classique, on peut souvent s'adapter aux vitesses minimax standards, par exemple grâce à des méthodes de sélection de modèles, cf. Massart [2007]. L'adaptation se fait parfois à des facteurs logarithmiques près, et les cas où l'adaptation est impossible sont plus exotiques, cf. Cai [2012]. En bandits stochastiques, la pénalisation de l'exploration fait émerger des phénomènes tout à fait différents dans les problèmes d'adaptation, comme nous avons pu le voir dans cette introduction.

Resituons les exemples précédemment discutés avec ces nouvelles notations.

Exemple 1.3 (Adaptation au support). *Pour l'adaptation au support, on prend* $\mathcal{A} = [K]$ *et les paramètres sont les bornes des supports des paiements :* $\Theta = \big\{(m, M) \in \mathbb{R}^2 : \quad m < M\big\}$ *; les classes d'environnements considérées sont* $\mathcal{E}(\kappa) = \mathcal{E}_{[m,M]}$.

La vitesse

$$B^\star(m, M, T) = (M - m)\sqrt{KT} \quad \text{pour } \mathcal{E}_{[m,M]}$$

est une vitesse minimax optimale à support connu d'après (1.28). *Et en utilisant la procéeure décrite au Chapitre 4, Section 4.5, on montre qu'on peut s'adapter aux vitesses minimax classiques, et donc qu'on peut atteindre la vitesse adaptative*

$$B^{\text{ADA}}(m, M, T) = B^\star(m, M, T) = (M - m)\sqrt{KT}(\Big(\sqrt{\ln K}\Big)) \,.$$

Exemple 1.4 (Adaptation à la régularité). *Dans le cas de l'adaptation à la régularité, on prend* $\mathcal{A} = [0, 1]$, *et le paramètre est l'exposant de Hölder* α, *qui parcourt l'ensemble* $\Theta = \{\alpha > 0\}$. *Les environnements sont donc les* $\mathcal{E}(\kappa) = \mathcal{E}^{\text{H}}(\alpha)$ *définis dans la Section 1.4.*

La vitesse

$$B^\star(\alpha, T) = T^{(\alpha+1)/(2\alpha+1)} \quad \text{pour } \mathcal{E}^{\text{H}}(\alpha)$$

est une vitesse minimax dans le cas de bandits continus (à paiements dans $[0, 1]$*). Cette vitesse* $B^\star(\alpha, T)$ *est inatteignable si l'on ne connait pas* α. *En revanche, l'algorithme* `MeDZO` *nous permet d'obtenir, entre autres, la vitesse adaptative*

$$B^{\text{ADA}}(\alpha, T) = (\ln T)^{3/2}\, T^{(\alpha+2)/(2\alpha+2)} \,,$$

et cette vitesse n'est (aux facteurs logarithmiques près) pas améliorable.

1.5.2. Adaptation et bandits linéaires ?

En fait, n'importe quel problème de bandits où une hypothèse est faite sur la classe d'environnements donne potentiellement naissance à des questions d'adaptation, dès que l'on souhaite se passer de l'hypothèse. Parmi ces problèmes, les bandits linéaires apparaissent comme une riche source d'inspiration.

Bandits linéaires. Les bandits linéaires forment une classe importante de problèmes de bandits (consulter [Lattimore and Szepesvári, 2020, Partie V] pour une description exhaustive) ; nous présentons rapidement le cadre, et soutenons qu'il s'agit d'un lieu naturel pour les problématiques d'adaptation. Soit d un entier naturel. On considère un ensemble d'actions

$$\mathcal{A} \subset \mathbb{R}^d \,,$$

où chaque action $\vec{a} \in \mathcal{A}$ est donc un vecteur. Dans les problèmes de bandits linéaires, le paiement moyen dépend linéairement de l'action choisie ; on suppose ainsi qu'il existe un vecteur $\vec{\mu} \in \mathbb{R}^d$ tel que

$$\mathrm{E}(\nu_{\vec{a}}) = \big\langle \vec{a},\, \vec{\mu} \big\rangle \,.$$

L'ensemble dans lequel on restreint le vecteur de paiements-moyens est déterminant dans la définition de la classe d'environnements en question. Prenons pour l'exemple $\mathcal{B}(R)$ la boule ℓ_2 de rayon R centrée en $(0, \dots, 0)$; fixons aussi pour l'exemple un bruit gaussien de variance 1. La classe d'environnements est alors

$$\mathcal{E}^{\text{lin}}\big(\mathcal{B}(R)\big) = \big\{(\nu_{\vec{a}}) : \quad \text{il existe } \vec{\mu} \in \mathcal{B}(R) \text{ pour tout } \vec{a} \in \mathcal{A} \quad \nu_{\vec{a}} = \mathcal{N}\big(\big\langle \vec{a},\, \vec{\mu} \big\rangle, 1\big)\big\} \,.$$

On peut alors, comme dans les exemples discutés précédemment, dresser les vitesses minimax à R fixé (celles-ci dépendront aussi de l'ensemble d'actions $\mathcal{A}$). L'étape suivante consistera à poursuivre l'étude sans la connaissance du rayon R.

Chapter 2.

KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints

Abstract

In the context of K–armed stochastic bandits with distribution only assumed to be supported by $[0, 1]$, we introduce the first algorithm, called KL-UCB-switch, that enjoys *simultaneously* a distribution-free regret bound of optimal order $\sqrt{KT}$ and a distribution-dependent regret bound of optimal order as well, that is, matching the $\kappa \ln T$ lower bound by Lai and Robbins [1985] and Burnetas and Katehakis [1996]. This self-contained contribution simultaneously presents state-of-the-art techniques for regret minimization in bandit models, and an elementary construction of non-asymptotic confidence bounds based on the empirical likelihood method for bounded distributions.

This work was led in collaboration with Aurélien Garivier, Pierre Ménard and Gilles Stoltz. The preprint, Garivier et al. [2018] is currently under review.

Contents

2.1. Introduction and brief literature review

Great progress has been made, over the last decades, in the understanding of the stochastic K–armed bandit problem. In this simplistic and yet paradigmatic sequential decision model, an agent can at each step $t \in \mathbb{N}^*$ sample one out of K independent sources of randomness and receive the corresponding outcome as a reward. The most investigated challenge is to minimize the regret, which is defined as the difference between the cumulated rewards obtained by the agent and by an oracle knowing in hindsight the distribution with largest expectation.

After Thompson's seminal paper (Thompson, 1933) and Gittins' Bayesian approach in the 1960s, Lai and his co-authors wrote in the 1980s a series of articles laying the foundations of a frequentist analysis of bandit strategies based on confidence regions. Lai and Robbins [1985] provided a general asymptotic lower bound, for parametric bandit models: for any reasonable strategy, the regret after T steps grows at least as $\kappa \ln(T)$, where κ is an informational complexity measure of the problem. In the 1990s, Agrawal [1995b] and Burnetas and Katehakis [1996] analyzed the UCB algorithm (see also the later analysis by Auer et al., 2002a), a simple procedure where at step t the arm with highest upper confidence bound is chosen. The same authors also extended the lower bound by Lai and Robbins to non-parametric models.

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, 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 and have a regret not much larger than $\kappa \ln(T)$) and distribution-free optimality (the maximal regret over all considered probability distributions should be of the optimal order $\sqrt{KT}$).

These efforts were pursued by further works in those three directions. Maillard et al. [2011] and Garivier and Cappé [2011] simultaneously proved that the distribution-dependent lower bound could be reached with exactly the right multiplicative constant in simple settings (for example, for binary rewards) and provided finite-time bounds to do so. They were followed by similar results

for other index policies like BayesUCB (Kaufmann et al., 2012) or Thompson sampling (Korda et al., 2013).

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 successfully 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 (distribution-free) optimality was proved, but distribution-dependent optimality was then not considered. It took a few more years before Ménard and Garivier [2017] and 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 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 2.1) and from a distribution-dependent viewpoint in the considered class of distributions (Theorem 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 $-\ln(\ln(T))$ in the distribution-dependent bound (Theorem 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. And according to the numerical experiments (see Section 2.3) IMED indeed does not seem to be optimal from a distribution-free viewpoint.

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 2.4 and 2.5).

Organization of the chapter. Section 2.2 presents the main contributions of this chapter: a description 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 2.3 discusses some numerical experiments comparing the performance of an empirically tuned version of the KL-UCB-switch algorithm to competitors like IMED or KL-UCB. The focus is not only set on the growth of the regret with time, but also on its dependency with respect to the number K of arms. Section 2.4 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 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), though sometimes in appendix, which makes this chapter fully self-contained. These known results are used as building blocks in Section 2.5 and 2.6, where the main results of this chapter are proved: Section 2.5 is devoted to distribution-free bounds, while Section 2.6 focuses on distribution-dependent bounds. An appendix provides the proofs of the classical material presented in Section 2.4, whenever these proofs did not fit in a few lines: anytime analysis of the MOSS strategy (Appendix 2.A) and proofs of the regularity and deviation results on the $\mathcal{K}_{\inf}$ quantity mentioned above (Appendix 2.B), which might be of independent interest. It also features the proof of a sophisticated distribution-dependent regret bound in the case of a known T: a regret bound with an optimal second order term (Appendix 2.C).

2.2. Setting and statement of the main results

We consider the simplest case of a bounded stochastic bandit problem, with finitely many arms indexed by $a \in \{1, \dots, K\}$ and with rewards in $[0,1]$. We denote by $\mathcal{P}[0,1]$ the set of probability distributions over $[0,1]$: each arm a is associated with an unknown probability distribution $\nu_a \in \mathcal{P}[0,1]$. 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 reward 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 sub-optimal 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 sub-optimal 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 $K \geqslant 2$,

$$\sup_{\underline{\nu}} R_T \geqslant \frac{1}{20} \min\Big\{\sqrt{KT},\, T\Big\} \,, \tag{2.1}$$

where the supremum is taken over all bandit problems $\underline{\nu}$ over $[0,1]$. Hence, a strategy is called optimal from a distribution-free viewpoint if there exists a numerical constant C such that fall $K \geqslant 2$, for all bandit problems $\underline{\nu}$ over $[0,1]$, for all $T \geqslant 1$, the regret is bounded by $R_T \leqslant C\sqrt{KT}$.

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. We recall its definition: consider two probability distributions ν, ν' over $[0,1]$. We write $\nu \ll \nu'$ when ν is absolutely continuous with respect to ν', and denote by $\mathrm{d}\nu/\mathrm{d}\nu'$ the density (the Radon-Nikodym derivative) of ν with respect to ν'. Then,

$$\mathrm{KL}(\nu, \nu') = \begin{cases} \displaystyle\int_{[0,1]} \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right) \mathrm{d}\nu & \text{if } \nu \ll \nu'; \\ +\infty & \text{otherwise.} \end{cases}$$

Now, the key information-theoretic quantity for stochastic bandit problems is given by an infimum of Kullback-Leibler divergences: 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 $+\infty$. Because of this convention, we may equivalently define $\mathcal{K}_{\inf}$ as

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) : \quad \nu'_a \in \mathcal{P}[0,1] \quad \text{with} \quad \nu_a \ll \nu'_a \quad \text{and} \quad \mathrm{E}(\nu'_a) > x\Big\}. \tag{2.2}$$

As essentially proved by Lai and Robbins [1985] and Burnetas and Katehakis [1996]—see also Garivier et al., 2019—, for any "reasonable" strategy, for any bandit problem $\underline{\nu}$ over $[0,1]$, for any sub-optimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,. \tag{2.3}$$

A strategy is called optimal from a distribution-dependent viewpoint if the reverse inequality holds with a lim sup instead of a lim inf, for any bandit problem $\underline{\nu}$ over $[0,1]$ and for any sub-optimal arm a.

By a "reasonable" strategy above, 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 sub-optimal arms a,

$$\forall\, \alpha > 0, \qquad \mathbb{E}\big[N_a(T)\big] = \mathrm{o}(T^\alpha)\,;$$

there exist such strategies, for instance, the UCB strategy already mentioned above. For uniformly super-fast convergent strategies, that is, strategies for which there actually exists a constant C such for all bandit problems $\underline{\nu}$ over $[0,1]$, for all sub-optimal arms a,

$$\frac{\mathbb{E}\big[N_a(T)\big]}{\ln T} \leqslant \frac{C}{\Delta_a^2}$$

(again, UCB is such a strategy), the lower bound above can be strengthened into: for any bandit problem $\underline{\nu}$ over $[0,1]$, for any sub-optimal arm a,

$$\mathbb{E}\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - \Omega\big(\ln(\ln T)\big)\,, \tag{2.4}$$

see Garivier et al. [2019, Section 4]. This order of magnitude $-\ln(\ln T)$ 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].

2.2.1. The KL-UCB-switch algorithm

Algorithm 2.1 Generic index policy

Inputs: `index functions` U_a
Initialization: `Play each arm` $a = 1, \dots, K$ `once and compute the` $U_a(K)$
for $t = K, \dots, T-1$ **do**
  `Pull an arm` $A_{t+1} \in \operatorname*{argmax}_{a=1,\dots,K} U_a(t)$
  `Get a reward` Y_{t+1} `drawn independently at random according to` $\nu_{A_{t+1}}$
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

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\!\left(\frac{T}{KN_a(t)}\right)}\,, \tag{2.5}$$

where $\ln_+$ denotes the non-negative part of the natural logarithm, $\ln_+ = \max\{\ln, 0\}$.

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

$$U_a^{\mathrm{KL}}(t) \stackrel{\mathrm{def}}{=} \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)} \ln_+\!\left(\frac{T}{KN_a(t)}\right) \right\}. \tag{2.6}$$

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. The purpose is to capture the good properties of both algorithms. In the sequel we will take $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$. More precisely, we define the index functions

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

The reasons for the choice of a threshold $f(T,K) = \lfloor (T/K)^{1/5} \rfloor$ will become clear in the proof of Theorem 2.1. Note that asymptotically KL-UCB-switch should behave like KL-UCB–type algorithm, as for large T we expect the number of pulls of a sub-optimal arm to be of order $N_a(t) \sim \ln(T)$ and optimal arms to have been played linearly many times, entailing $U_a^{\mathrm{M}}(t) \approx U_a^{\mathrm{KL}}(t) \approx \widehat{\mu}_a(t)$.

Since we are considering distributions over $[0,1]$, the data-processing inequality for Kullback-Leibler divergences ensures (see, e.g., Garivier et al., 2019, Lemma 1) that for all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in \big(\mathrm{E}(\nu), 1\big)$,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \inf_{\nu' : \mathrm{E}(\nu') > \mu} \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}\big(\mathrm{E}(\nu')\big)\Big) = \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}(\mu)\Big)\,,$$

where $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. Therefore, by Pinsker's inequality for Bernoulli distributions,

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant 2\big(\mathrm{E}(\nu) - \mu\big)^2\,, \qquad \text{thus} \qquad U_a^{\mathrm{KL}}(t) \leqslant U_a^{\mathrm{M}}(t) \tag{2.7}$$

for all arms a and all rounds $t \geqslant K$. In particular, this actually shows that KL-UCB-switch interpolates between KL-UCB and MOSS,

$$U_a^{\mathrm{KL}}(t) \leqslant U_a(t) \leqslant U_a^{\mathrm{M}}(t)\,. \tag{2.8}$$

2.2.2. 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 two theorems below are provided in Sections 2.5 and 2.6, respectively.

Theorem 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*

$$R_T \leqslant (K-1) + 23\sqrt{KT}\,.$$

KL-UCB-switch thus enjoys a distribution-free regret bound of optimal order $\sqrt{KT}$, see (2.1). The MOSS strategy by Audibert and Bubeck [2009] already enjoyed this optimal distribution-free regret bound but its construction (relying on a sub-Gaussian assumption) prevents it from being optimal from a distribution-dependent viewpoint.

Theorem 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,*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T\big((\ln T)^{2/3}\big)\,,$$

where a finite-time, closed-form expression of the $\mathcal{O}_T\big((\ln T)^{2/3}\big)$ term is given by (2.39) for the choice $\delta = (\ln T)^{-1/3}$.

By considering the exact same algorithm but by following a more sophisticated proof we may in fact get a stronger result, whose (extremely technical) proof is deferred to Appendix 2.C.

Theorem 2.3 (Distribution-dependent bound with a second-order term). *We actually have, when $\mu^\star \in (0,1)$ and $T \geqslant K/(1-\mu^\star)$,*

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

where a finite-time, closed-form expression of the $\mathcal{O}_T(1)$ term is provided in (2.57).

KL-UCB-switch thus enjoys a distribution-dependent regret bounds of optimal orders, see (2.3) and (2.4). This optimal order was already reached by 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 $[0,1]$.

2.2.3. 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 $\mathcal{O}\big(\sqrt{KT}\big)$ and $\big(1+\mathrm{o}(1)\big)(\ln T)/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$ bounds. Indeed, there are first, two different rates, $\sqrt{T}$ and $\ln T$, 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 (which is non-decreasing) by

$$\varphi(x) = \ln_+\big(x(1+\ln_+^2 x)\big) \tag{2.9}$$

and the associated index functions by

$$U_a^{\text{KL-A}}(t) \stackrel{\text{def}}{=} \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)}\, \varphi\left(\frac{t}{K N_a(t)}\right) \right\} \tag{2.10}$$

$$U_a^{\text{M-A}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\, \varphi\left(\frac{t}{K N_a(t)}\right)}\,. \tag{2.11}$$

A careful comparison of (2.10) and (2.11) to (2.5) and (2.6) shows that $U_a^{\text{KL-A}}(t) \leqslant U_a^{\text{KL}}(t)$ and

$$U_a^{\text{M-A}}(t) \leqslant U_a^{\text{M},\varphi}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\, \varphi\left(\frac{T}{K N_a(t)}\right)} \tag{2.12}$$

when all these quantities are based on the same past (i.e., when they are defined for the same algorithm).

The -A in the superscripts stands for "augmented" or for "anytime" as this augmented exploration gives rise to the anytime version of KL-UCB-switch, which simply relies on the index

$$U_a^{\text{A}}(t) = \begin{cases} U_a^{\text{KL-A}}(t) & \text{if } N_a(t) \leqslant f(t,K) \\ U_a^{\text{M-A}}(t) & \text{if } N_a(t) > f(t,K) \end{cases} \tag{2.13}$$

where $f(T,K) = \lfloor (t/K)^{1/5} \rfloor$. Note that the thresholds $f(t,K)$ when the switches occur from the sub-index $U_a^{\text{KL-A}}(t)$ to the other sub-index $U_a^{\text{M-A}}(t)$ now vary with t (and we cannot exclude that a switch back may occur).

For this anytime version of KL-UCB-switch, the same ranking of (sub-)indexes holds as the one (2.8) for our first version of KL-UCB-switch relying on the horizon T:

$$U_a^{\text{KL-A}}(t) \leqslant U_a^{\text{A}}(t) \leqslant U_a^{\text{M-A}}(t)\,. \tag{2.14}$$

The performance guarantees are indicated in the next two theorems, whose proofs may be found in Sections 2.5 and 2.6, respectively. 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 $\ln_+$ as the exploration function.

Theorem 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 $T \geqslant 1$,*

$$R_T \leqslant (K-1) + 44\sqrt{KT}\,.$$

Theorem 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, for all $T \geqslant 1$,*

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \mathcal{O}_T\big((\ln T)^{6/7}\big)$$

where a finite-time, closed-form expression of the $\mathcal{O}_T\big((\ln T)^{6/7}\big)$ term is given by Equation (2.32) for the choice $\delta = (\ln T)^{-1/7}$.

2.3. Numerical experiments

We provide here some numerical experiments comparing the different algorithms we refer to in this work. The KL-UCB-switch, KL-UCB, and MOSS algorithms are used in their anytime versions as described in Section 2.2.1 and Section 2.2.3. However, we stick to the natural exploration function $\ln_+\big(t/(KN_a(t))\big)$, i.e., without extra-exploration. For KL-UCB-switch we actually consider a slightly delayed switch function, different from the one in our theoretical analysis: $f(t,K) = \lfloor t/K \rfloor^{8/9}$, which generally exhibits a good empirical performance. While our choice $f(t,K) = \lfloor t/K \rfloor^{1/5}$ appeared to be a good choice for minimizing the theoretical upper bounds, many other choices (such as the one considered in the experiments below) would also have been possible, at the cost of larger constants in one of the two regret bounds.

Distribution-dependent bounds. We compare in Figure 2.1 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 shape 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 lies between the one of MOSS and the one of KL-UCB.

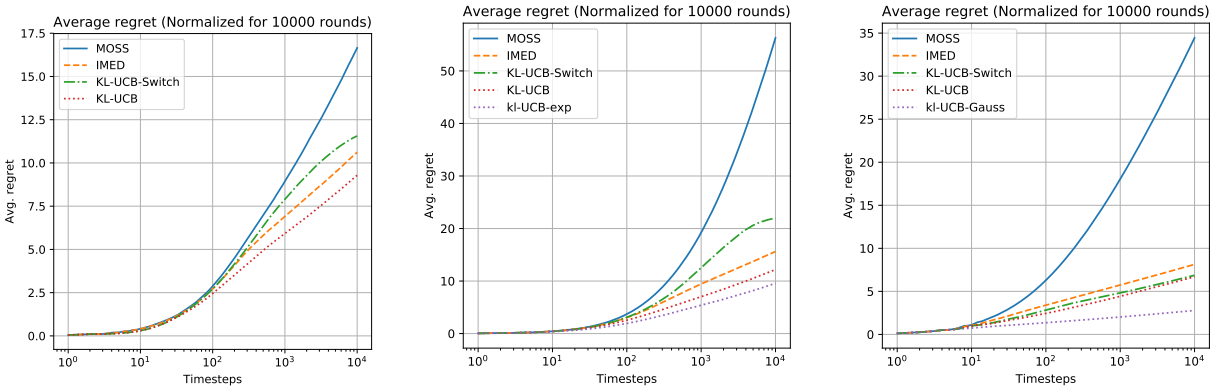


Figure 2.1.: Regrets approximated over $10,000$ runs, shown on a logarithmic scale; distributions of the arms consist of:
Left: Bernoulli distributions with parameters (0.9, 0.8)
Middle: Exponential distributions with expectations (0.15, 0.12, 0.10, 0.05), truncated on $[0,1]$,
Right: Gaussian distributions with means (0.7, 0.5, 0.3, 0.2) and same standard deviation $\sigma = 0.1$, truncated on $[0,1]$

Distribution-free bounds. Here we also consider the UCB algorithm of Auer et al. [2002a] with the exploration function $\ln(t)$. We plot the behavior of the normalized regret, $R_T/\sqrt{KT}$, either as a function of T (Figure 2.2 left) or of K (Figure 2.2 right). This quantity should remain bounded as T or K increases. KL-UCB-switch and KL-UCB have a normalized regret that does 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 regrets of UCB and IMED seem to suffer from a sub-optimal dependence in K.

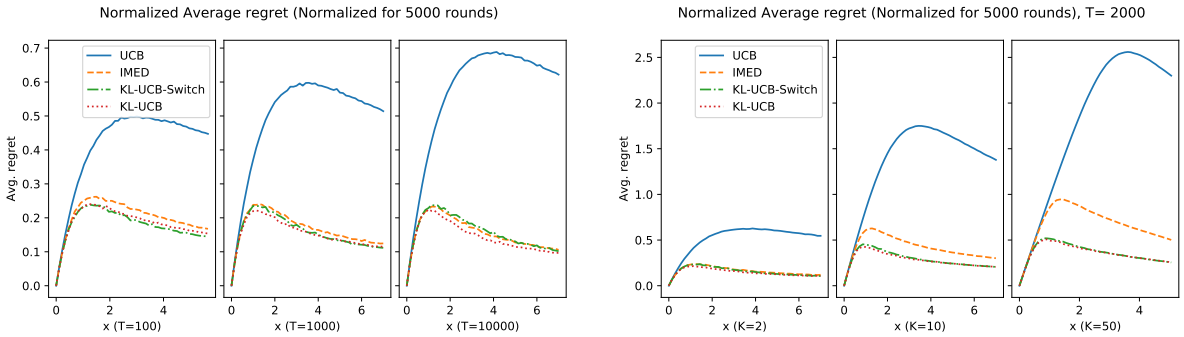


Figure 2.2.: Expected regret $R_T/\sqrt{KT}$, approximated over $5,000$ runs
Left: as a function of x, for a Bernoulli bandit problem with parameters $\big(0.8,\ 0.8 - x\sqrt{K/T}\big)$ and for time horizons $T \in \{100,\ 1000,\ 10000\}$
Right: as a function of x, for a Bernoulli bandit problem with parameters $(0.8,\ 0.8 - x\sqrt{K/T},\ \ldots,\ 0.8 - x\sqrt{K/T})$ and K arms, where $K \in \{2,\ 10,\ 50\}$

2.4. Results (more or less) 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). *Readers familiar with the material described here are urged to move to the next section.*

2.4.1. Optional skipping—how to go from global times t to local times n

The trick detailed here is standard in the bandit literature, see, e.g., its application in Auer et al. [2002a]. It is sometimes called optional skipping, and sometimes, optional sampling; we pick the first terminology, following what seems to be the preferred terminology in probability theory[1]. In any case, the original reference is Theorem 5.2 of Doob [1953, Chapter III, p. 145]; one can also check Chow and Teicher [1988, Section 5.3] for a more recent reference.

Doob's optional skipping enables the rewriting of 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 $N_a(t) = n$ that the specific arm considered has been pulled so far. 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. Now, Doob's optional skipping 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.

Example 1 (Controlling an empirical average). *Recall that $N_a(t) \geqslant 1$ for $t \geqslant K$ and $N_a(t) \leqslant t-K+1$ 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,

$$\mathbb{P}\big[\widehat{\mu}_a(t) \in \mathcal{E}\big] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E} \ \text{ and } \ N_a(t) = n\big] \leqslant \sum_{n=1}^{t-K+1} \mathbb{P}\big[\widehat{\mu}_{a,n} \in \mathcal{E}\big]\,.$$

[1] The abstract of a recent article by Simons et al. [2002] reads: "A general set of distribution-free conditions is described under which an i.i.d. sequence of random variables is preserved under optional skipping. This work is motivated by theorems of J.L. Doob (1936) and Z. Ignatov (1977), unifying and extending aspects of both."

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.

Actually, a more careful use of optional skipping would be

$$\mathbb{P}\big[\widehat{\mu}_a(t) \in \mathcal{E}\big] \leqslant \mathbb{P}\left[\bigcup_{n=1}^{t-K+1} \big\{\widehat{\mu}_{a,n} \in \mathcal{E}\big\}\right] = \mathbb{P}\Big[\exists\, n \in \{1,\dots,t-K+1\} :\ \widehat{\mu}_{a,n} \in \mathcal{E}\Big]\,.$$

2.4.2. Maximal version of Hoeffding's inequality

The maximal version of Hoeffding's inequality (Proposition 2.1) is a standard result from Hoeffding [1963]. It was already used in the original analysis of MOSS (Audibert and Bubeck, 2009). For our slightly simplified analysis of MOSS (see Section 2.4.3), we will rather rely on Corollary 2.1, a consequence of Proposition 2.1 obtained by integrating it.

Proposition 2.1. *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 $u \geqslant 0$ and for all $N \geqslant 1$:*

$$\mathbb{P}\left[\max_{n \geqslant N} \big(\widehat{\mu}_n - \mu\big) \geqslant u\right] \leqslant \mathrm{e}^{-2Nu^2}\,. \tag{2.15}$$

Corollary 2.1. *Under the same assumptions, for all $\varepsilon \geqslant 0$,*

$$\mathbb{E}\left[\Big(\max_{n \geqslant N} \big(\mu - \widehat{\mu}_n - \varepsilon\big)\Big)^+\right] \leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2}\,. \tag{2.16}$$

Of course, Proposition 2.1 and Corollary 2.1 hold by symmetry with $\mu - \widehat{\mu}_n$ instead of $\widehat{\mu}_n - \mu$.

Proof. By the Fubini-Tonelli theorem, an integration of the maximal deviation inequality (2.15) yields

$$\begin{aligned}\mathbb{E}\left[\Big(\max_{n \geqslant N} \big(\mu - \widehat{\mu}_n - \varepsilon\big)\Big)^+\right] &= \int_0^{+\infty} \mathbb{P}\Big[\max_{n \geqslant N} \big(\widehat{\mu}_n - \mu - \varepsilon\big) \geqslant u\Big]\mathrm{d}u \\ &\leqslant \int_0^{+\infty} \mathrm{e}^{-2N(u+\varepsilon)^2}\mathrm{d}u \leqslant \mathrm{e}^{-2N\varepsilon^2}\int_0^{+\infty} \mathrm{e}^{-2Nu^2}\mathrm{d}u = \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{N}}\,\mathrm{e}^{-2N\varepsilon^2}\,.\end{aligned}$$

□

2.4.3. Distribution-free bound for the MOSS algorithm

Such a distribution-free bound was already provided 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 only provide a slightly shorter and more focused proof of these results based on Corollary 2.1 and indicate an intermediate result—see (2.17)—that will be useful for us in the analysis of our new KL-UCB-switch algorithm. We do not claim any improvement on the results themselves, just a clarification of the existing proofs.

Our proof is slightly shorter and more focused for two reasons. First, in the two references mentioned, the peeling trick was used on the probabilities of deviations (see Proposition 2.1) 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, based on Corollary 2.1 and at fixed times depending solely on T. Second, unlike the two mentioned references, we do not attempt to simultaneously build a distribution-free and some type of distribution-dependent

bound. This raised technical difficulties because of the correlations between the choices of the arms and the observed rewards. The idea of our approach is to focus solely on the distribution-free regime, for which we notice that some crude bounding neglecting the correlations suffice (i.e., our analysis deals with all sub-optimal arms in the same way, independently of how often they are played).

For a known horizon T, we denote by A^{M}_{t+1} the arm played by the index strategy maximizing, at each step $t+1$ with $t \geqslant K$, the quantities (2.5):

$$U^{\mathrm{M}}_a(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\left(\frac{T}{KN_a(t)}\right)}\,.$$

The superscripts M in A^{M}_{t+1} and $U^{\mathrm{M}}_a(t)$ stand for MOSS. We do so not to mix it with the arm A_{t+1} played by the KL-UCB-switch strategy (no superscript), but of course, once an arm a was sufficiently pulled, we have $A_{t+1} = A^{\mathrm{M}}_{t+1}$ by definition of the KL-UCB-switch strategy.

Appendix 2.A provides the proof of the following regret bound. We denote by $a^\star$ an optimal arm, i.e., an arm such that $\mu_a = \mu^\star$.

Proposition 2.2. *For a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, MOSS achieves a regret bound smaller than $R_T \leqslant (K-1) + 17\sqrt{KT}$. More precisely, with the notation of optional skipping (Section 2.4.1), we have the inequalities*

$$\begin{aligned} R_T &= T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A^{\mathrm{M}}_t}\right] \\ &\leqslant (K-1) + \underbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t-1)\big)^+\Big]}_{\leqslant 13\sqrt{KT}} \\ &\quad + \underbrace{\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} + \sqrt{\frac{\ln_+\big(T/(Kn)\big)}{2n}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]}_{\leqslant 4\sqrt{KT}} \end{aligned} \tag{2.17}$$

Remark 1. *The proof (see Remark 4) actually reveals that for a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, and for all strategies (not only MOSS), the following bound holds:*

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{M}}_{a^\star}(t-1)\big)^+\Big] \leqslant 13\sqrt{KT}\,.$$

We will re-use this fact to state a similar remark below (Remark 2), which will be useful for Part 2 of the proof lying in Section 2.5.

Our proof in Appendix 2.A reveals that designing an adaptive version of MOSS comes at no effort. For this adaptive version we will also want to possibly explore more. We will do so by considering an augmented exploration function φ, that is, a function $\varphi \geqslant \ln_+$ as in (2.9). We therefore define MOSS-anytime (M-A) as relying on the indexes defined in (2.11), which we copy here:

$$U^{\text{M-A}}_a(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \varphi\left(\frac{t}{KN_a(t)}\right)}\,.$$

We denote by $A^{\text{M-A}}_{t+1}$ the arm picked as $\operatorname*{argmax}_{a=1,\dots,K} U^{\text{M-A}}_a(t)$.

Proposition 2.3. *For all horizons $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, MOSS-anytime achieves a regret bound smaller than $R_T \leqslant (K-1) + c\sqrt{KT}$ where $c = 30$ for $\varphi = \ln_+$ and $c = 33$ for the augmented exploration function $\varphi(x) = \ln_+\big(x(1+\ln_+^2 x)\big)$ defined in (2.9). More precisely, with the notation of optional skipping (Section 2.4.1), we have the inequalities*

$$\begin{aligned} R_T &= T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t^{\text{M-A}}}\right] \\ &\leqslant (K-1) + \underbrace{\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big]}_{\leqslant 26\sqrt{KT}} \\ &\quad + \underbrace{\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(\widehat{\mu}_{a,n} + \sqrt{\frac{\varphi\big(T/(Kn)\big)}{2n}} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right]}_{\leqslant 4\sqrt{KT} \text{ for } \varphi=\ln_+ \text{ and } 7\sqrt{KT} \text{ for } \varphi(x)=\ln_+((x(1+\ln_+^2 x))} \end{aligned} \tag{2.18}$$

Remark 2. *Similarly to above, the proof (see Remark 4) actually reveals that for a known horizon $T \geqslant 1$, for all bandit problems $\underline{\nu}$ over $[0,1]$, and for all strategies (not only MOSS-anytime), the following bound holds:*

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star}^{\text{M-A}}(t-1)\big)^+\Big] \leqslant 26\sqrt{KT}\,.$$

This remark will be useful for Part 2 of the proof lying in Section 2.5.

2.4.4. Regularity and deviation/concentration results on $\mathcal{K}_{\inf}$

We start with a quantification of the (left-)regularity of $\mathcal{K}_{\inf}$ and then provide a deviation and a concentration result on $\mathcal{K}_{\inf}$.

Regularity of $\mathcal{K}_{\inf}$

The lower left-semi-continuity (2.19) first appeared as Lemma 7 in Honda and Takemura [2015], see also Garivier et al. [2019, Lemma 3] for a later but simpler proof. The upper left-semi-continuity (2.20) relies on the same arguments as (2.7), namely, the data-processing inequality for Kullback-Leibler divergences and Pinsker's inequality. These two inequalities are proved in detail in Appendix 2.B; the proposed proofs are slightly simpler or lead to sharper bounds than in the mentioned references.

Lemma 2.1 (regularity of $\mathcal{K}_{\inf}$). *For all $\nu \in \mathcal{P}[0,1]$ and all $\mu \in (0,1)$,*

$$\forall \varepsilon \in [0,\mu]\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + \frac{\varepsilon}{1-\mu}\,, \tag{2.19}$$

and

$$\forall \varepsilon \in \big[0, \mu - \mathrm{E}(\nu)\big]\,, \qquad \mathcal{K}_{\inf}(\nu,\mu) \geqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + 2\varepsilon^2\,. \tag{2.20}$$

We draw two consequences from Lemma 2.1: the left-continuity of $\mathcal{K}_{\inf}$ and a useful inclusion in terms of level sets.

Corollary 2.2. *For all $\nu \in \mathcal{P}[0,1]$, the function $\mathcal{K}_{\inf}(\nu, \cdot) : \mu \in (0,1) \mapsto \mathcal{K}_{\inf}(\nu,\mu)$ is left-continuous. In particular, on the one hand, $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)\big) = 0$ whenever $\mathrm{E}(\nu) \in (0,1)$, and on the other hand, for all $\nu \in \mathcal{P}[0,1]$ and $\mu \in (0,1)$,*

$$\mathcal{K}_{\inf}(\nu,\mu) = \inf\Big\{\mathrm{KL}(\nu,\nu') : \quad \nu' \in \mathcal{P}[0,1] \quad \text{and} \quad \mathrm{E}(\nu') \geqslant \mu\Big\}.$$

Proof. The left-continuity follows from a sandwich argument via the upper bound (2.19) and the lower bound $\mathcal{K}_{\inf}(\nu,\mu-\varepsilon) \leqslant \mathcal{K}_{\inf}(\nu,\mu)$ that holds for all $\varepsilon \in [0,\mu]$ by the very definition of $\mathcal{K}_{\inf}$. The fact that $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)-\varepsilon\big) = 0$ for all $\varepsilon \in \big(0, \mathrm{E}(\nu)\big]$ thus entails, in particular, that $\mathcal{K}_{\inf}\big(\nu, \mathrm{E}(\nu)\big) = 0$. □

Corollary 2.3. *For all $\nu \in \mathcal{P}[0,1]$, all $\mu \in (0,1)$, all $u > 0$, and all $\varepsilon > 0$,*

$$\big\{\mathcal{K}_{\inf}(\nu,\mu-\varepsilon) > u\big\} \subseteq \big\{\mathcal{K}_{\inf}(\nu,\mu) > u + 2\varepsilon^2\big\}.$$

Proof. We apply (2.20) and merely need to explain why the condition $\varepsilon \in \big[0, \mu - \mathrm{E}(\nu)\big]$ therein is satisfied. Indeed, $\mathcal{K}_{\inf}(\nu,\mu-\varepsilon) > u > 0$ indicates in particular that $\mu - \varepsilon > \mathrm{E}(\nu)$, or put differently, $\varepsilon < \mu - \mathrm{E}(\nu)$. □

Deviation results on $\mathcal{K}_{\inf}$

We provide two deviation results on $\mathcal{K}_{\inf}$: first, in terms of probabilities of deviations and next, in terms of expected deviations.

The first deviation inequality was essentially provided by Cappé et al. [2013, Lemma 6]. For the sake of completeness, we recall its proof in Section 2.B.

Proposition 2.4 (deviation result on $\mathcal{K}_{\inf}$)**.** *Let $\widehat{\nu}_n$ denote the empirical distribution associated with a sequence of $n \geqslant 1$ i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geqslant 0$,*

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

A useful corollary in terms of expected deviations can now be stated.

Corollary 2.4 (integrated deviations for $\mathcal{K}_{\inf}$)**.** *Under the same assumptions, for all $\varepsilon > 0$, the index*

$$U_{\varepsilon,n} = \sup\Big\{\mu \in [0,1] \;\Big|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_n,\mu\big) \leqslant \varepsilon\Big\}$$

satisfies

$$\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] \leqslant (2n+1)\,\mathrm{e}^{-n\varepsilon}\sqrt{\frac{\pi}{n}}$$

Proof. By the Fubini-Tonelli theorem, just as in the proof of Corollary 2.1 (for the first two equalities), and subsequently using the definition of $U_{\varepsilon,n}$ as a supremum (for the third equality, together with the left-continuity of $\mathcal{K}_{\inf}$ deriving from Lemma 2.1), we have

$$\begin{aligned}\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] &= \int_0^{+\infty} \mathbb{P}\Big[\mathrm{E}(\nu) - U_{\varepsilon,n} > u\Big]\mathrm{d}u = \int_0^{+\infty} \mathbb{P}\Big[U_{\varepsilon,n} < \mathrm{E}(\nu) - u\Big]\mathrm{d}u \\ &= \int_0^{+\infty} \mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)-u\big) > \varepsilon\Big]\mathrm{d}u\,.\end{aligned}$$

Now, Corollary 2.3 (for the first inequality) and the deviation inequality of Proposition 2.4 (for the second inequality) indicate that for all $u > 0$,

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

Combining all elements, we get

$$\mathbb{E}\Big[\big(\mathrm{E}(\nu) - U_{\varepsilon,n}\big)^+\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n\varepsilon} \int_0^{+\infty} \mathrm{e}^{-2nu^2}\mathrm{d}u = \mathrm{e}(2n+1)\,\mathrm{e}^{-n\varepsilon}\,\frac{1}{2}\sqrt{\frac{\pi}{2n}}\,.$$

from which the stated bound follows, as $\mathrm{e}/\big(2\sqrt{2}\big) \leqslant 1$. □

Concentration result on $\mathcal{K}_{\mathrm{inf}}$

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 2.B.

Proposition 2.5 (concentration result on $\mathcal{K}_{\mathrm{inf}}$)**.** *With the same notation and assumptions as in the previous proposition, consider a real number $\mu \in \big(\mathrm{E}(\nu), 1\big)$ and define*

$$\gamma = \frac{1}{\sqrt{1-\mu}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu}\right)\right). \tag{2.21}$$

Then for all $x < \mathcal{K}_{\mathrm{inf}}(\nu,\mu)$,

$$\mathbb{P}\big[\mathcal{K}_{\mathrm{inf}}(\widehat{\nu}_n,\mu) \leqslant x\big] \leqslant \begin{cases} \exp(-n\gamma/8) \leqslant \exp(-n/4) & \text{if } x \leqslant \mathcal{K}_{\mathrm{inf}}(\nu,\mu) - \gamma/2 \\ \exp\Big(-n\big(\mathcal{K}_{\mathrm{inf}}(\nu,\mu) - x\big)^2/(2\gamma)\Big) & \text{if } x > \mathcal{K}_{\mathrm{inf}}(\nu,\mu) - \gamma/2 \end{cases}.$$

2.5. Proofs of the distribution-free bounds: Theorems 2.1 and 2.4

The two proofs are extremely similar; we show, for instance, Theorem 2.4 and explain how to adapt the proof for Theorem 2.1. The first steps of the proof(s) use the exact same arguments as in the proofs of the performance bounds of MOSS (Propositions 2.2 and 2.3, see Appendix 2.A) in the exact same order. We explain below why we had to copy them and had to resort to the intermediary bounds for MOSS stated in the indicated propositions.

We recall that we denote by $a^\star$ an optimal arm, i.e., an arm such that $\mu_a = \mu^\star$. We first apply a trick introduced by Bubeck and Liu [2013]: by definition of the index policy, for $t \geqslant K$,

$$U^{\mathrm{A}}_{a^\star}(t) \leqslant \max_{a=1,\dots,K} U^{\mathrm{A}}_a(t) = U^{\mathrm{A}}_{A_{t+1}}(t)$$

so that the regret of KL-UCB-switch is bounded by

$$R_T = \sum_{t=1}^{T} \mathbb{E}\big[\mu^\star - \mu_{A_t}\big] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U^{\mathrm{A}}_{a^\star}(t-1)\big] + \sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big]\,. \quad (2.22)$$

Part 1: We first deal with the second sum in (2.22) and successively use $x \leqslant \delta + (x-\delta)^+$ for all x and δ for the first inequality; the fact that $U^{\mathrm{A}}_a(t) \leqslant U^{\text{M-A}}_a(t) \leqslant U^{\mathrm{M},\varphi}_a(t)$ by (2.12) and (2.14), for the second inequality; and optional skipping (Section 2.4.1) for the third inequality, keeping in mind that pairs (a,n) such $A_t = a$ and $N_a(t-1) = n$ correspond to at most one round $t \in \{K+1,\dots,T\}$:

$$\begin{aligned}
\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big] &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right] \\
&\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{A_t}(t-1) - \mu_{A_t} - \sqrt{\frac{K}{T}}\right)^+\right] \quad (2.23) \\
&\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{a,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right] \quad (2.24)
\end{aligned}$$

where we recall that

$$U^{\mathrm{M},\varphi}_{a,n} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}\,.$$

We now apply one of the bounds of Proposition 2.3 to further bound the sum at hand by

$$\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\mathrm{A}}_{A_t}(t-1) - \mu_{A_t}\big] \leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{M},\varphi}_{a,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right] \leqslant 7\sqrt{KT}\,.$$

Remark 3. *We may now explain why we copied the beginning of the proof of Proposition 2.3 and why we cannot just say that the ranking $U^{\mathrm{A}}_a(t) \leqslant U^{\text{M-A}}_a(t)$ entails that the regret of the anytime version of KL-UCB-switch is bounded by the regret of the anytime version of MOSS. Indeed, it is difficult to relate*

$$\sum_{t=K+1}^{T} \mathbb{E}\big[U^{\text{M-A}}_{A_t}(t-1) - \mu_{A_t}\big] \qquad \text{and} \qquad \sum_{t=K+1}^{T} \mathbb{E}\big[U^{\text{M-A}}_{A^{\text{M-A}}_t}(t-1) - \mu_{A^{\text{M-A}}_t}\big]$$

as the two series of arms A_t (picked by KL-UCB-switch) and $A^{\text{M-A}}_t$ (picked by the adaptive version of MOSS) cannot be related. Hence, it is difficult to directly bound quantities like (2.23)*. However,*

the proof of the performance bound of MOSS relies on optional skipping and considers, in some sense, all possible values a for the arms picked: it controls the quantity (2.24)*, which appears as a regret bound that is achieved by all index policies with indexes smaller than the ones of the anytime version of MOSS.*

Part 2: We now deal with the first sum in (2.22). We take positive parts, get back to the definition (2.13) of $U^{\mathrm{A}}_{a^\star}(t-1)$, and add some extra non-negative terms:

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbb{E}\big[\mu^\star - U^{\mathrm{A}}_{a^\star}(t-1)\big] \leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{A}}_{a^\star}(t-1)\big)^+\Big] \\
&\leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{M-A}}_{a^\star}(t-1)\big)^+ \underbrace{\mathbb{1}_{\{N_{a^\star}(t-1)> f(t-1,K)\}}}_{\leqslant 1}\Big] \\
&\leqslant \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{M-A}}_{a^\star}(t-1)\big)^+\Big].
\end{aligned}$$

Now, the bound (2.18) of Proposition 2.3, together with the Remark 2, indicates that

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{M-A}}_{a^\star}(t-1)\big)^+\Big] \leqslant 26\sqrt{KT}\,.$$

Note that Remark 2 exactly explains that for the sum above we do not bump into the issues raised in Remark 3 for the other sum in (2.22).

Part 3: Integrated deviations in terms of $\mathcal{K}_{\inf}$ divergence. We showed so far that the distribution-free regret bound of the anytime version of KL-UCB-switch was given by the (intermediary) regret bound (2.18) of Proposition 2.3, which is smaller than $(K-1)+33\sqrt{KT}$, plus

$$\begin{aligned}
\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1)\leqslant f(t-1,K)\}}\Big] &= \sum_{t=K}^{T-1} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)\leqslant f(t,K)\}}\Big] \\
&\leqslant \sum_{t=K}^{T-1} \sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star,t,n}\big)^+\Big] \qquad (2.25)
\end{aligned}$$

where we applied optional skipping (Section 2.4.1) and where we denoted

$$U^{\text{KL-A}}_{a^\star,t,n} = \sup\left\{\mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n},\mu\big) \leqslant \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right\} \qquad (2.26)$$

the counterpart of the quantity $U^{\text{KL-A}}_{a^\star}(t)$ defined in (2.10). Here, the additional subscript t in $U^{\text{KL-A}}_{a^\star,t,n}$ refers to the denominator of $t/(Kn)$ in the $\varphi(t/(Kn))$ term.

Now, Corollary 2.4 exactly indicates that for each given t and all $n \geqslant 1$,

$$\mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star,t,n}\big)^+\Big] \leqslant (2n+1)\sqrt{\frac{\pi}{n}}\,\exp\left(-\varphi\Big(\frac{t}{Kn}\Big)\right).$$

The t considered are such that $t \geqslant K$ and thus, $f(t,K) \leqslant (t/K)^{1/5} \leqslant t/K$. Therefore, the considered n are such that $1 \leqslant n \leqslant f(t,K)$ and thus, $t/(Kn) \geqslant 1$. Given that $\varphi \geqslant \ln_+$, we proved

$$\mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star,t,n}\big)^+\Big] \leqslant (2n+1)\sqrt{\frac{\pi}{n}}\,\frac{Kn}{t} = \frac{K\sqrt{\pi}}{t}\,(2n+1)\sqrt{n}\,.$$

We sum this bound over $n \in \{1, \dots, f(t/K)\}$, using again that $f(t,K) \leqslant (t/K)^{1/5}$:

$$\sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star,t,n}\big)^+\Big] \leqslant \frac{K\sqrt{\pi}}{t} \sum_{n=1}^{f(t,K)} \underbrace{(2n+1)\sqrt{n}}_{\leqslant 3f(t,K)^{3/2}} \leqslant \frac{3K\sqrt{\pi}}{t} \underbrace{f(t,K)^{5/2}}_{\leqslant (t/K)^{1/2}} \leqslant 3\sqrt{\pi}\sqrt{\frac{K}{t}}\,.$$

We substitute this inequality into (2.25):

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star}(t-1)\big)^+ \mathbb{1}_{\{N_{a^\star}(t-1) \leqslant f(t-1,K)\}}\Big] \\ \leqslant \sum_{t=K}^{T-1} \sum_{n=1}^{f(t,K)} \mathbb{E}\Big[\big(\mu^\star - U^{\text{KL-A}}_{a^\star,t,n}\big)^+\Big] \leqslant 3\sqrt{\pi} \underbrace{\sum_{t=K}^{T-1} \sqrt{\frac{K}{t}}}_{\leqslant 2\sqrt{KT}, \text{ see (2.40)}} \leqslant 6\sqrt{\pi}\sqrt{KT} \leqslant 11\sqrt{KT}\,.\end{aligned}$$

The final regret bound is obtained as the sum of this $11\sqrt{KT}$ bound plus the $(K-1) + 33\sqrt{KT}$ bound obtained above. This concludes the proof of Theorem 2.4.

Part 4: Adaptations needed for Theorem 2.1, i.e., to analyze the version of KL-UCB-switch relying on the knowledge of the horizon T. Parts 1 and 2 of the proof remain essentially unchanged, up to the (intermediary) regret bound to be applied now: (2.17) of Proposition 2.2, which is smaller than $(K-1) + 17\sqrt{KT}$. The additional regret bound, accounting, as we did in Part 3, for the use of KL-UCB-indexes for small T, is no larger than

$$\begin{aligned}\sum_{t=K}^{T-1} \sum_{n=1}^{f(T,K)} (2n+1)\sqrt{\frac{\pi}{n}} \exp\left(-\ln_+\left(\frac{T}{Kn}\right)\right) &= \sum_{t=K}^{T-1} \sum_{n=1}^{f(T,K)} (2n+1)\sqrt{\frac{\pi}{n}}\frac{Kn}{T} \\ &= K\sqrt{\pi} \sum_{n=1}^{f(T,K)} \underbrace{(2n+1)\sqrt{n}}_{\leqslant 3f(T,K)^{3/2}} \leqslant 3\sqrt{\pi}\,K\,f(T,K)^{5/2} \leqslant 3\sqrt{\pi}\,K\sqrt{\frac{T}{K}} \leqslant 6\sqrt{KT}\,.\end{aligned}$$

This yields the claimed $(K-1) + 23\sqrt{KT}$ bound.

2.6. Proofs of the distribution-dependent bounds: Theorems 2.2 and 2.5

The proofs below can be adapted (simplified) to provide an elementary analysis of performance of the KL-UCB algorithm on the class of all distributions over a bounded interval, by keeping only Parts 1 and 2 of the proofs below. The study of KL-UCB in Cappé et al. [2013] remained somewhat intricate and limited to finitely supported distributions.

We provide first an anytime analysis, i.e., the proof of Theorem 2.5, and then explain the simplifications in the analysis (and improvements in the second-order terms in the regret bound) arising when the horizon T is known, i.e., as far as the proof of Theorem 2.2 is concerned.

2.6.1. Proof of Theorem 2.5

The proof starts as in Cappé et al. [2013]. We fix a sub-optimal arm a. Given $\delta \in (0, \mu^\star)$ sufficiently small (to be determined by the analysis), we first decompose $\mathbb{E}\big[N_a(T)\big]$ as

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] &= 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[A_{t+1} = a\big] \\ &= 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) < \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] .\end{aligned}$$

We then use that by definition of the index policy, $A_{t+1} = a$ only if $U_a^{\mathrm{A}}(t) \geqslant U_{a^\star}^{\mathrm{A}}(t)$, where we recall that $a^\star$ denotes an optimal arm (i.e., an arm such that $\mu_a = \mu^\star$). We also use $U_{a^\star}^{\mathrm{A}}(t) \geqslant U_{a^\star}^{\text{KL-A}}(t)$, which was stated in (2.14). We get

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] &\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\mathrm{A}}(t) < \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] \\ &\leqslant 1 + \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a\big] .\end{aligned}$$

Finally, by the definition (2.13) of $U_a^{\mathrm{A}}(t)$, we proved so far

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant 1 &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) > f(t,K)\big] . \qquad (2.27)\end{aligned}$$

We now deal with each of the three sums above.

Part 1: We first deal with the first sum in (2.27) and to that end, fix some $t \in \{K, \dots, T-1\}$. By the definition (2.10) of $U_{a^\star}^{\text{KL-A}}(t)$ as a supremum,

$$\mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \leqslant \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star}(t), \mu^\star - \delta\big) > \frac{1}{N_{a^\star}(t)}\, \varphi\left(\frac{t}{K N_{a^\star}(t)}\right)\right] .$$

By optional skipping (see Section 2.4.1), applied with some care,

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star}(t), \mu^\star - \delta\big) > \frac{1}{N_{a^\star}(t)}\,\varphi\Big(\frac{t}{KN_{a^\star}(t)}\Big)\right] \leqslant \mathbb{P}\left[\exists n \in \{1,\dots,t-K+1\} : \;\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right].$$

Now, for $n \geqslant \lfloor t/K \rfloor + 1$ and given the definition (2.9) of φ, we have $\varphi\big(t/(Kn)\big) = 0$. By definition, $\mathcal{K}_{\inf}(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta) > 0$ requires in particular that the expectation $\widehat{\mu}_{a^\star,n}$ of $\widehat{\nu}_{a^\star,n}$ be smaller than $\mu^\star - \delta$. This fact, together with a union bound, implies

$$\begin{aligned}
&\mathbb{P}\left[\exists n \in \{1,\dots,t-K+1\} : \;\; \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right] \\
&\leqslant \mathbb{P}\Big[\exists n \geqslant \lfloor t/K \rfloor + 1 : \;\; \widehat{\mu}_{a^\star,n} \leqslant \mu^\star - \delta\Big] + \sum_{n=1}^{\lfloor t/K \rfloor} \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right].
\end{aligned}$$

Hoeffding's maximal inequality (Proposition 2.1) upper bounds the first term by $\exp(-2\delta^2 t/K)$, while Corollary 2.3 and Proposition 2.4 provide the upper bound

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n}\,\varphi\Big(\frac{t}{Kn}\Big)\right] \leqslant \mathrm{e}(2n+1)\,\exp\Big(-n\Big(2\delta^2 + \varphi\big(t/(Kn)\big)/n\Big)\Big).$$

Collecting all inequalities, we showed so far that

$$\mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] \leqslant \exp(-2\delta^2 t/K) + \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big).$$

Summing over $t \in \{K, \dots, T-1\}$, using the formula for geometric series, on the one hand, and performing some straightforward (and uninteresting) calculation detailed below in Lemma 2.2 on the other hand, we finally bound the first sum in (2.27) by

$$\begin{aligned}
\sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\text{KL-A}}(t) < \mu^\star - \delta\big] &\leqslant \sum_{t=K}^{T-1} \exp(-2\delta^2 t/K) + \sum_{t=K}^{T-1} \sum_{n=1}^{\lfloor t/K \rfloor} \mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2 - \varphi\big(t/(Kn)\big)\Big) \\
&\leqslant \frac{1}{1-\mathrm{e}^{-2\delta^2/K}} + \frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}\,.
\end{aligned}$$

This concludes the first part of this proof.

Part 2: We then deal with the second sum in (2.27). We introduce

$$\widetilde{U}_a^{\text{KL-A}}(t) \stackrel{\text{def}}{=} \sup\left\{\mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\big(\widehat{\nu}_a(t), \mu\big) \leqslant \frac{1}{N_a(t)}\,\varphi\Big(\frac{T}{KN_a(t)}\Big)\right\}$$

that only differs from the original index $U_a^{\text{KL-A}}(t)$ defined in (2.10) by the replacement of $t/(Kn)$ by $T/(Kn)$ as the argument of φ. Therefore, we have $\widetilde{U}_a^{\text{KL-A}}(t) \geqslant U_a^{\text{KL-A}}(t)$. Replacing also $f(t,K)$

by the larger quantity $f(T,K)$, the second sum in (2.27) is therefore bounded by

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big] \\
&\leqslant \sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(T,K)\Big] \qquad (2.28)\\
&\leqslant \sum_{n=1}^{f(T,K)} \sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big] .
\end{aligned}$$

Optional skipping (see Section 2.4.1) indicates that for each value of n,

$$\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}\Big[\widetilde{U}_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big] \\
&\qquad\qquad = \sum_{t=K}^{T-1} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big]
\end{aligned}$$

where $U_{a^\star,T,n}^{\text{KL-A}}$ was defined in (2.26). We now observe that the events $\big\{A_{t+1} = a \ \text{ and } \ N_a(t) = n\big\}$ are disjoint as t varies in $\{K,\dots,T-1\}$. Therefore,

$$\sum_{t=K}^{T-1} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) = n\Big] \leqslant \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big] .$$

All in all, we proved so far that

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{KL-A}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(t,K)\big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big] . \qquad (2.29)$$

Now, note that the supremum in (2.26) is taken over a closed interval, as $\mathcal{K}_{\inf}$ is non-decreasing in its second argument (by its definition as an infimum) and as $\mathcal{K}_{\inf}$ is left-continuous (Corollary 2.2). This supremum is therefore a maximum. Hence, by distinguishing the cases where $U_{a^\star,T,n}^{\text{KL-A}} = \mu^\star - \delta$ and $U_{a^\star,T,n}^{\text{KL-A}} > \mu^\star - \delta$, we have the equality of events

$$\Big\{U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big\} = \left\{\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big)\right\} .$$

We assume that $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \frac{1-\mu^\star}{2}\, \mathcal{K}_{\inf}(\nu_a, \mu^\star)$$

to hold and introduce

$$n_1 = \left\lceil \frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \right\rceil \geqslant 1 .$$

For $n \geqslant n_1$, by definition of n_1,

$$\frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big) \leqslant \underbrace{\frac{\varphi\big(T/(Kn)\big)}{\varphi(T/K)}}_{\leqslant 1} \left(\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \frac{2\delta}{1-\mu^\star}\right) \leqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star) - \frac{2\delta}{1-\mu^\star}$$

while by the regularity property (2.19), we have $\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \geqslant \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) - \frac{\delta}{1-\mu^\star}$. We therefore proved that for $n \geqslant n_1$,

$$\begin{aligned}\mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big] &= \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n}\varphi\Big(\frac{T}{Kn}\Big)\right] \\ &\leqslant \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{\delta}{1-\mu^\star}\right].\end{aligned}$$

Therefore we may resort to the concentration inequality on $\mathcal{K}_{\inf}$ stated as Proposition 2.5. We set $x = \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)$ and simply sum the bounds obtained in the two regimes considered therein:

$$\mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) - \frac{\delta}{1-\mu^\star}\right] \leqslant \mathrm{e}^{-n/4} + \exp\Big(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\Big)$$

where $\gamma_\star$ was defined in (2.21). For $n \leqslant n_1 - 1$, we bound the probability at hand by 1. Combining all these arguments together yields

$$\begin{aligned}\sum_{n=1}^{f(T,K)} \mathbb{P}\Big[U_{a^\star,T,n}^{\text{KL-A}} \geqslant \mu^\star - \delta\Big] &\leqslant n_1 - 1 + \sum_{n=n_1}^{f(T,K)} \mathrm{e}^{-n/4} + \sum_{n=n_1}^{f(T,K)} \exp\Big(-\frac{n\delta^2}{2\gamma_\star(1-\mu^\star)^2}\Big) \\ &\leqslant \frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} + \underbrace{\frac{1}{1-\mathrm{e}^{-1/4}}}_{\leqslant 5} + \underbrace{\frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}}_{=\mathcal{O}(1/\delta^2)}\end{aligned}$$

where the second inequality follows from the formula for geometric series and from the definition of n_1.

Part 3: We then deal with the third sum in (2.27). This sum involves the indexes $U_a^{\text{M-A}}(t)$ only when $N_a(t) > f(t,K)$, that is, when $N_a(t) \geqslant f(t,K)+1$, where $f(t,K) = \lfloor (t/K)^{1/5} \rfloor$. Under the latter condition, the indexes are actually bounded by

$$U_a^{\text{M-A}}(t) \stackrel{\text{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{t}{KN_a(t)}\Big)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2(t/K)^{1/5}}\,\varphi\big((t/K)^{4/5}\big)}}_{\to 0 \text{ as } t\to\infty}.$$

We denote by $T_0(\Delta_a, K)$ the smallest time T_0 such that for all $t \geqslant T_0$,

$$\sqrt{\frac{1}{2(t/K)^{1/5}}\,\varphi\big((t/K)^{4/5}\big)} \leqslant \frac{\Delta_a}{4}. \tag{2.30}$$

This time T_0 only depends on K and Δ_a; a closed-form upper bound on its value could be easily provided. With this definition, we already have that the sum of interest may be bounded by

$$\begin{aligned}&\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(t,K)\big] \\ &\leqslant T_0(\Delta_a, K) + \sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) + \Delta_a/4 \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(t,K)\Big] \\ &\leqslant T_0(\Delta_a, K) + \sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/2 \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(t,K)\Big]\end{aligned}$$

where for the second inequality, we assumed that $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \frac{\Delta_a}{4}$$

to hold. Optional skipping (see Section 2.4.1), using that the events $\big\{A_{t+1} = a \text{ and } N_a(t) = n\big\}$ are disjoint as t varies, as already done between (2.28) and (2.29), provides the upper bound

$$\begin{aligned}\sum_{t=T_0(\Delta_a,K)}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/2 \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(t,K)\Big] \\ \leqslant \sum_{n\geqslant 1} \mathbb{P}\big[\widehat{\mu}_{a,n} \geqslant \mu_a + \Delta_a/2\big] \leqslant \sum_{n\geqslant 1} \mathrm{e}^{-n\Delta_a^2/2} = \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}}\end{aligned}$$

where the second inequality is due to Hoeffding's inequality (in its non-maximal version, see Proposition 2.1). A summary of the bound thus provided in this part is:

$$\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\text{M-A}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(t,K)\big] \leqslant T_0(\Delta_a, K) + \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}} = \mathcal{O}(1)$$

where $T_0(\Delta_a, K)$ was defined in (2.30).

Part 4: Conclusion of the proof of Theorem 2.5. Collecting all previous bounds and conditions, we proved that when $\delta \in (0, \mu^\star)$ is sufficiently small for

$$\delta < \min\left\{\frac{1-\mu^\star}{2}\,\mathcal{K}_{\inf}(\nu_a, \mu^\star),\ \frac{\Delta_a}{4}\right\} \tag{2.31}$$

to hold, then

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \\ + \underbrace{\frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}}_{=\mathcal{O}(1/\delta^6)} + \underbrace{\frac{1}{1-\mathrm{e}^{-2\delta^2/K}} + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}}_{=\mathcal{O}(1/\delta^2)} + \underbrace{T_0(\Delta_a,K) + \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}} + 6}_{=\mathcal{O}(1)}\end{aligned} \tag{2.32}$$

where

$$\frac{\varphi(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} = \frac{\ln T + \ln\ln T + \mathcal{O}(1)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} = \frac{\ln T + \ln\ln T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} + \mathcal{O}(\delta \ln T)\,.$$

The leading term in this regret bound is $\ln T/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$, while the order of magnitude of the smaller-order terms is given by

$$\delta \ln T + \frac{1}{\delta^6} = \mathcal{O}\big((\ln T)^{6/7}\big)$$

for δ of the order of $(\ln T)^{-1/7}$. When T is sufficiently large, this value of δ is smaller than the required threshold (2.31).

It only remains to state and prove Lemma 2.2 (used at the very end of the first part of the proof above).

Lemma 2.2. *We have the bound*

$$\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K\rfloor}\mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2-\varphi\big(t/(Kn)\big)\Big)\leqslant\frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}\,.$$

Proof. The double sum can be rewritten, by permuting the order of summations, as

$$\begin{aligned}\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K\rfloor}\mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2-\varphi\big(t/(Kn)\big)\Big)&=\sum_{n=1}^{\lfloor T/K\rfloor}\sum_{t=Kn}^{T-1}\mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2-\varphi\big(t/(Kn)\big)\Big)\\&=\sum_{n=1}^{\lfloor T/K\rfloor}\mathrm{e}(2n+1)\,\exp\big(-2n\delta^2\big)\sum_{t=Kn}^{T-1}\exp\Big(-\varphi\big(t/(Kn)\big)\Big)\,.\end{aligned}$$

We first fix $n\geqslant 1$ and use that $t\mapsto\exp\big(-\varphi(t/(Kn)\big)$ is non-increasing to get

$$\sum_{t=Kn}^{T-1}\exp\Big(-\varphi\big(t/(Kn)\big)\Big)\leqslant 1+\int_{Kn}^{T-1}\exp\Big(-\varphi\big(t/(Kn)\big)\Big)\mathrm{d}t=1+Kn\int_1^{(T-1)/(Kn)}\exp\big(-\varphi(u)\big)\mathrm{d}u$$

where we operated the change of variable $u=t/(Kn)$. Now, by the change of variable $v=\ln(u)$,

$$\begin{aligned}\int_1^{(T-1)/(Kn)}\exp\big(-\varphi(u)\big)\mathrm{d}u\leqslant\int_1^{+\infty}\exp\big(-\varphi(u)\big)\mathrm{d}u&=\int_1^{+\infty}\frac{1}{u\big(1+\ln^2(u)\big)}\mathrm{d}u\\&=\int_0^{+\infty}\frac{1}{1+v^2}\mathrm{d}v=\big[\arctan\big]_0^{+\infty}=\frac{\pi}{2}\,.\end{aligned}$$

All in all, we proved so far that

$$\begin{aligned}\sum_{t=K}^{T-1}\sum_{n=1}^{\lfloor t/K\rfloor}\mathrm{e}(2n+1)\,\exp\Big(-2n\delta^2-\varphi\big(t/(Kn)\big)\Big)&\leqslant\sum_{n=1}^{\lfloor T/K\rfloor}\mathrm{e}(2n+1)\big(1+Kn\pi/2\big)\,\exp\big(-2n\delta^2\big)\\&\leqslant\sum_{n=1}^{+\infty}\mathrm{e}\big(1+(2+K\pi/2)n+K\pi n^2\big)\,\exp\big(-2n\delta^2\big)\,.\end{aligned}$$

To conclude our calculation, we use that by differentiation of series, for all $\theta>0$,

$$\sum_{m=0}^{+\infty}\mathrm{e}^{-m\theta}=\frac{1}{1-\mathrm{e}^{-\theta}}$$

$$-\sum_{m=1}^{+\infty}m\,\mathrm{e}^{-m\theta}=\frac{-\mathrm{e}^{-\theta}}{(1-\mathrm{e}^{-\theta})^2}\qquad\text{thus}\qquad\sum_{m=1}^{+\infty}m\,\mathrm{e}^{-m\theta}\leqslant\frac{1}{(1-\mathrm{e}^{-\theta})^2}\tag{2.33}$$

$$\sum_{m=1}^{+\infty}m^2\,\mathrm{e}^{-m\theta}=\frac{\mathrm{e}^{-\theta}(1+\mathrm{e}^{-\theta})}{(1-\mathrm{e}^{-\theta})^3}\leqslant\frac{2}{(1-\mathrm{e}^{-\theta})^3}\,.\tag{2.34}$$

Hence, taking $\theta=2\delta^2$,

$$\sum_{n=1}^{+\infty}\mathrm{e}\big(1+(2+K\pi/2)n+K\pi n^2\big)\,\exp\big(-2n\delta^2\big)\leqslant\frac{\mathrm{e}}{1-\mathrm{e}^{-2\delta^2}}+\frac{\mathrm{e}(2+K\pi/2)}{(1-\mathrm{e}^{-2\delta^2})^2}+\frac{2\mathrm{e}K\pi}{(1-\mathrm{e}^{-2\delta^2})^3}\leqslant\frac{\mathrm{e}(3+8K)}{(1-\mathrm{e}^{-2\delta^2})^3}$$

which concludes the proof of this lemma. □

2.6.2. Proof of Theorem 2.2

We adapt (simplify) the proof of Theorem 2.5, by replacing the thresholds $f(t,K)$ by $f(T,K)$, by taking $\varphi = \ln_+$, etc. To that end, we start with a similar decomposition,

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant 1 &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\mathrm{KL}}(t) < \mu^\star - \delta\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant f(T,K)\big] \\ &+ \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) > f(T,K)\big]\,. \end{aligned} \tag{2.35}$$

The first sum is bounded using exactly the same arguments as in the proof of Theorem 2.5 (optional skipping, Hoeffding's maximal inequality, Corollary 2.3 and Proposition 2.4): for all $t \in \{K, \dots, T-1\}$,

$$\begin{aligned}&\mathbb{P}\big[U_{a^\star}^{\mathrm{KL}}(t) < \mu^\star - \delta\big] \\ &\leqslant \mathbb{P}\Big[\exists n \geqslant \lfloor T/K \rfloor + 1 : \ \widehat{\mu}_{a^\star,n} \leqslant \mu^\star - \delta\Big] + \sum_{n=1}^{\lfloor T/K \rfloor} \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{1}{n} \ln\left(\frac{T}{Kn}\right)\right] \\ &\leqslant \exp(-2\delta^2 T/K) + \mathrm{e}(2n+1)\, \exp\Big(-n\Big(2\delta^2 + \ln\big(T/(Kn)\big)/n\Big)\Big) \\ &= \exp(-2\delta^2 T/K) + \frac{\mathrm{e}K}{T}\,(2n^2+n)\exp(-2n\delta^2)\,. \end{aligned}$$

Summing over t and substituting the bounds (2.33)–(2.34), we proved

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}^{\mathrm{KL}}(t) < \mu^\star - \delta\big] &\leqslant T\exp(-2\delta^2 T/K) + \frac{\mathrm{e}K}{T}\left(\frac{4}{(1-\mathrm{e}^{-2\delta^2})^3} + \frac{1}{(1-\mathrm{e}^{-2\delta^2})^2}\right) \\ &\leqslant T\exp(-2\delta^2 T/K) + \frac{5\mathrm{e}K}{T(1-\mathrm{e}^{-2\delta^2})^3}\end{aligned}$$

For the second sum in (2.35), we note that the initial manipulations in Part 2 of the proof of Theorem 2.5 are unnecessary in the case of Theorem 2.2; we may directly start at (2.28) and the rest of the arguments used and calculation performed then hold word for word, under the same condition that $\delta < (1-\mu^\star)\mathcal{K}_{\inf}(\nu_a, \mu^\star)/2$. We get, with the same notation,

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant f(T,K)\big] \\ \leqslant \frac{\ln(T/K)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 2\delta/(1-\mu^\star)} + 5 + \frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}\,.\end{aligned} \tag{2.36}$$

The third sum in (2.35) involves the indexes $U_a^{\mathrm{M}}(t)$ only under the condition $N_a(t) > f(T,K)$, in which case $N_a(t) \geqslant (T/K)^{1/5}$ and

$$U_a^{\mathrm{M}}(t) \stackrel{\mathrm{def}}{=} \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\ln_+\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \sqrt{\frac{1}{2\,(T/K)^{1/5}}\,\ln_+\big((T/K)^{4/5}\big)}\,.$$

We mimic the proof scheme of Part 3 of the proof of Theorem 2.5 and start by assuming that T is sufficiently large for

$$\sqrt{\frac{1}{2\,(T/K)^{1/5}}\,\ln_+\big((T/K)^{4/5}\big)} \leqslant \frac{\Delta_a}{4} \tag{2.37}$$

to hold. Under the same condition $\delta < \Delta_a/4$, we get, by a careful application of optional skipping using that the events $\big\{A_{t+1} = a \;\text{ and }\; N_a(t) = n\big\}$ are disjoint as t varies and by Hoeffding's inequality,

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \;\text{ and }\; A_{t+1} = a \;\text{ and }\; N_a(t) > f(T,K)\big] \\ \leqslant \sum_{n=f(T,K)+1}^{T-1} \mathbb{P}\big[\widehat{\mu}_{a,n} \geqslant \mu_a + \Delta_a/2\big] \leqslant \sum_{n \geqslant f(T,K)+1} \mathrm{e}^{-n\Delta_a^2/2} \leqslant \frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}}\,.\end{aligned} \tag{2.38}$$

Collecting all bounds, we proved that whenever T is sufficiently large for (2.37) to hold and whenever δ is sufficiently small for (2.31) to hold,

$$\begin{aligned}\mathbb{E}\big[N_a(T)\big] \leqslant \frac{\ln(T/K)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - 2\delta/(1-\mu^\star)} \\ + \underbrace{\frac{1}{1-\mathrm{e}^{-\delta^2/(2\gamma_\star(1-\mu^\star)^2)}}}_{=\mathcal{O}(1/\delta^2)} + \underbrace{\frac{1}{1-\mathrm{e}^{-\Delta_a^2/2}}}_{=\mathcal{O}(1)} + 6 + T\exp(-2\delta^2 T/K) + \underbrace{\frac{5\mathrm{e}K}{T(1-\mathrm{e}^{-2\delta^2})^3}}_{=\mathcal{O}(1/(T\delta^6))}\,.\end{aligned} \tag{2.39}$$

The leading term in this regret bound is $\ln T/\mathcal{K}_{\inf}(\nu_a,\mu^\star)$, while the order of magnitude of the smaller-order terms is given by

$$\delta\ln T + \frac{1}{\delta^2} + T\exp(-2\delta^2 T/K) + \frac{1}{T\delta^6} = \mathcal{O}\big((\ln T)^{2/3}\big)$$

for δ of the order of $(\ln T)^{-1/3}$. When T is sufficiently large, this value of δ is smaller than the required threshold (2.31).

Appendix for Chapter 2

Content and structure

2.A. A simplified proof of the regret bounds for MOSS(-anytime)

2.B. Proofs of the regularity and deviation/concentration results on $\mathcal{K}_{\inf}$

1. Proof of the regularity lemma (Lemma 2.1)
2. A useful tool: a variational formula for $\mathcal{K}_{\inf}$ (statement)
3. Proof of the deviation result (Proposition 2.4)
4. Proof of the concentration result (Proposition 2.5)

2.C. Proof of Theorem 2.3 (with the $-\ln\ln T$ term in the regret bound)

2.D. Proof of the variational formula (Lemma 2.3)

2.A. A simplified proof of the regret bounds for MOSS(-anytime)

This section provides the proofs of Propositions 2.2 and 2.3. To emphasize the similarity of the analyses in the anytime and non-anytime cases, 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 (2.5) and (2.11), which we both state with a generic exploration function φ. Indeed, compare

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{T}{KN_a(t)}\Big)} \qquad \text{and} \qquad U_a^{\text{M-A}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{t}{KN_a(t)}\Big)}\,.$$

We will denote by

$$U_{a,\tau}^{\mathrm{GM}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)}\,\varphi\Big(\frac{\tau}{KN_a(t)}\Big)}$$

the index of the 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 $(\tau_K, \dots, \tau_{T-1})$ of integers, either $\tau_t \equiv T$ for MOSS or $\tau_t = t$ for MOSS-anytime, and picks at each step $t+1$ with $t \geqslant K$, an arm A_{t+1}^{GM} with maximal index $U_{a,\tau_t}^{\mathrm{GM}}(t)$. For a given t, we denote by $U_{a,\tau_t,n}^{\mathrm{GM}}$ the quantities corresponding to $U_{a,\tau_t}^{\mathrm{GM}}(t)$ by optional skipping (see Section 2.4.1).

We provide below an analysis for increasing exploration functions $\varphi : (0,+\infty) \to [0,+\infty)$ such that φ vanishes on $(0,1]$ and $\varphi \geqslant \ln_+$, properties that are all satisfied for the two exploration functions stated in Proposition 2.3. The general result is stated as the next proposition.

Proposition 2.6. *For all bandit problems $\underline{\nu}$ over $[0,1]$, for all $T \geqslant 1$ and all sequences $(\tau_K, \dots, \tau_{T-1})$ bounded by T, the regret of the generic MOSS strategy described above, with an increasing exploration function $\varphi \geqslant \ln_+$ vanishing on $(0,1]$, is smaller than*

$$R_T \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_{t-1}}^{\mathrm{GM}}(t-1)\big)^+\Big] + \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big]$$

where

$$U_{a,T,n}^{\mathrm{GM}} = \widehat{\mu}_{a,n} + \sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}\,.$$

In addition,

$$\sum_{t=K+1}^{T} \mathbb{E}\Big[\big(\mu^\star - U_{a^\star,\tau_{t-1}}^{\mathrm{GM}}(t-1)\big)^+\Big] \leqslant \underbrace{20\sqrt{\frac{\pi}{8}}}_{\leqslant 12.6} \sum_{t=K}^{T-1} \sqrt{\frac{K}{\tau_t}}$$

and

$$\sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\Big[\big(U_{a,T,n}^{\mathrm{GM}} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sqrt{KT}\left(1 + \frac{\pi}{4} + \frac{1}{\sqrt{2}}\int_1^{+\infty} u^{-3/2}\sqrt{\varphi(u)}\mathrm{d}u\right).$$

The bounds of Propositions 2.2 and 2.3, including the intermediary bounds (2.17) and (2.18), follow from this general result, up to the following straightforward calculation. On the one hand, in the known horizon case $\sum 1/\sqrt{\tau_t} \leqslant T/\sqrt{T} = \sqrt{T}$, whereas in the anytime case,

$$\sum_{t=K}^{T-1} 1/\sqrt{\tau_t} = \sum_{t=K}^{T-1} 1/\sqrt{t} \leqslant \int_0^T \frac{1}{\sqrt{u}}\mathrm{d}u = 2\sqrt{T}\,. \tag{2.40}$$

On the other hand, by the change of variable $u = \mathrm{e}^{v^2}$,

$$\int_1^{+\infty} u^{-3/2}\sqrt{\ln(u)}\mathrm{d}u = 2\int_0^{+\infty} v^2\,\mathrm{e}^{-v^2/2}\mathrm{d}v = \sqrt{2\pi}$$

and, using well-known inequalities like $\sqrt{x+x'} \leqslant \sqrt{x} + \sqrt{x'}$ and $\ln(1+x) \leqslant x$ for $x, x' \geqslant 0$,

$$\begin{aligned}\int_1^{+\infty} \sqrt{u^{-3}\ln\bigl(u(1+\ln^2(u))\bigr)}\,\mathrm{d}u &\leqslant \int_1^{+\infty} \sqrt{u^{-3}\ln(u)}\,\mathrm{d}u + \int_1^{+\infty} \sqrt{u^{-3}\ln\bigl(1+\ln^2(u)\bigr)}\,\mathrm{d}u \\ &\leqslant \int_1^{+\infty} \sqrt{u^{-3}\ln(u)}\,\mathrm{d}u + \int_1^{+\infty} \sqrt{u^{-3}\ln^2(u)}\,\mathrm{d}u \\ &= 2\int_0^{+\infty} v^2\,\mathrm{e}^{-v^2/2}\mathrm{d}v + 2\int_0^{+\infty} v^3\,\mathrm{e}^{-v^2/2}\mathrm{d}v = \sqrt{2\pi} + 4\,.\end{aligned}$$

The constant 17 of Proposition 2.2 is obtained as an upper bound on the sum of $12.6 \leqslant 13$ and $1 + \pi/4 + \sqrt{\pi} \leqslant 3.6 \leqslant 4$. The constants 30 and 33 of Proposition 2.3 are respectively obtained as upper bounds on on the sum of $2 \times 12.6 \leqslant 26$ and $1 + \pi/4 + \sqrt{\pi} \leqslant 4$, and on the sum of $2 \times 12.6 \leqslant 26$ and $1 + \pi/4 + \sqrt{\pi} + 4/\sqrt{2} \leqslant 6.4 \leqslant 7$.

Proof. The beginning of this proof is completely similar to the beginning of the proof provided in Section 2.5.

The first step is standard, see Bubeck and Liu [2013]. By definition of the index policy, for $t \geqslant K$,

$$U^{\mathrm{GM}}_{a^\star,\tau_t}(t) \leqslant \max_{a=1,\dots,K} U^{\mathrm{GM}}_{a,\tau_t}(t) = U^{\mathrm{GM}}_{A^{\mathrm{GM}}_{t+1},\tau_t}(t)$$

so that the regret of the strategy is smaller than

$$R_T = \sum_{t=1}^{T} \mathbb{E}\bigl[\mu^\star - \mu_{A^{\mathrm{GM}}_t}\bigr] \leqslant (K-1) + \sum_{t=K+1}^{T} \mathbb{E}\bigl[\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\bigr] + \sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,\tau_{t-1}}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr]\,. \tag{2.41}$$

The term $K-1$ above accounts for the initial K rounds, when each arm is played once.

A preliminary transformation of the right-hand side of (2.41). We successively use the fact that the index $U^{\mathrm{GM}}_{a,\tau}(t-1)$ increases with τ since φ is increasing (for the first inequality below), $x \leqslant \delta + (x-\delta)^+$ for all x and δ (for the second inequality), and optional skipping (Section 2.4.1, for the third inequality), keeping in mind that pairs (a,n) such $A^{\mathrm{GM}}_t = a$ and $N_a(t-1) = n$ correspond to at most one round $t \in \{K+1,\dots,T\}$:

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,\tau_{t-1}}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr] &\leqslant \sum_{t=K+1}^{T} \mathbb{E}\Bigl[U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,T}(t-1) - \mu_{A^{\mathrm{GM}}_t}\Bigr] \\ &\leqslant \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{A^{\mathrm{GM}}_t,T}(t-1) - \mu_{A^{\mathrm{GM}}_t} - \sqrt{\frac{K}{T}}\right)^+\right] \\ &\leqslant \sqrt{KT} + \sum_{a=1}^{K}\sum_{n=1}^{T} \mathbb{E}\left[\left(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{\frac{K}{T}}\right)^+\right].\end{aligned}$$

While the last two inequalities may seem very crude, it turns out they are sharp enough to obtain the claimed distribution-free bounds. Moreover, they get rid of the bothersome dependencies

among the arms that are contained in the choice of the arms A_t^{GM}. Therefore, we have shown that the right-hand side of (2.41) is bounded by

$$\begin{aligned}&(K-1)+\sum_{t=K+1}^{T}\mathbb{E}\big[\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big]+\sum_{t=K+1}^{T}\mathbb{E}\Big[U^{\mathrm{GM}}_{A_t^{\mathrm{GM}},\tau_{t-1}}(t-1)-\mu_{A_t^{\mathrm{GM}}}\Big]\\ \leqslant&(K-1)+\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big)^+\Big]+\sqrt{KT}+\sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n}-\mu_a-\sqrt{K/T}\big)^+\Big]\,.\end{aligned} \tag{2.42}$$

This inequality actually holds for all choices of sequences $(\tau_t)_{K\leqslant t\leqslant T-1}$ with $\tau_t\leqslant T$. The first sum in the right-hand side of (2.42) depends on the specific value of $(\tau_t)_{K\leqslant t\leqslant T-1}$, and thus, on the specific MOSS algorithm considered, but the second sum only depends on T.

Control of the left deviations of the best arm, that is, of the first sum in (2.41) and (2.42). For each given round $t\in\{K,\dots,T-1\}$, 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 dealt with first by using $U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\geqslant\widehat{\mu}_{a^\star}(t)$, which actually holds with equality given $N_{a^\star}(t)\geqslant\tau_t/K$, and second, by optional skipping (Section 2.4.1) and by the integrated version of Hoeffding's inequality (Corollary 2.1):

$$\begin{aligned}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)\geqslant\tau_t/K\}}\Big]&\leqslant\mathbb{E}\Big[\big(\mu^\star-\widehat{\mu}_{a^\star}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)\geqslant\tau_t/K\}}\Big]\\&\leqslant\mathbb{E}\Big[\max_{n\geqslant\tau_t/K}\big(\mu^\star-\widehat{\mu}_{a^\star,n}\big)^+\Big]\leqslant\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,.\end{aligned} \tag{2.43}$$

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 $\big\{N_{a^\star}(t)<\tau_t/K\big\}$ along the geometric grid $x_\ell=\beta^{-\ell}\,\tau_t/K$, where $\ell=0,1,2,\dots$ (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):

$$\begin{aligned}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big]&=\sum_{\ell=0}^{+\infty}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+\mathbb{1}_{\{x_{\ell+1}\leqslant N_{a^\star}(t)<x_\ell\}}\Big]\\&\leqslant\sum_{\ell=0}^{+\infty}\mathbb{E}\Big[\max_{x_{\ell+1}\leqslant n<x_\ell}\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)^+\Big]\end{aligned}$$

where in the second inequality, we applied optional skipping (Section 2.4.1) once again. Now for any ℓ, the summand can be controlled as follows, first, by $\varphi\geqslant\ln_+=\ln$ on $[1,+\infty)$, second, by

using $n < x_\ell$ and third, by Corollary 2.1:

$$\begin{aligned}
\mathbb{E}\Big[\max_{x_{\ell+1}\leqslant n<x_\ell}\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t,n}\big)^+\Big] &= \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell}\left(\mu^\star - \widehat{\mu}_{a^\star,n} - \sqrt{\frac{1}{2n}\varphi\Big(\frac{\tau_t}{Kn}\Big)}\right)^+\right] \\
&\leqslant \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell}\left(\mu^\star - \widehat{\mu}_{a^\star,n} - \sqrt{\frac{1}{2n}\ln\Big(\frac{\tau_t}{Kn}\Big)}\right)^+\right] \\
&\leqslant \mathbb{E}\left[\max_{x_{\ell+1}\leqslant n<x_\ell}\left(\mu^\star - \widehat{\mu}_{a^\star,n} - \sqrt{\frac{1}{2x_\ell}\ln\Big(\frac{\tau_t}{Kx_\ell}\Big)}\right)^+\right] \\
&\leqslant \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\exp\left(-\frac{x_{\ell+1}}{x_\ell}\ln\Big(\frac{\tau_t}{Kx_\ell}\Big)\right) \\
&= \sqrt{\frac{\pi}{8}}\sqrt{\frac{1}{x_{\ell+1}}}\big(\beta^{-\ell}\big)^{1/\beta} = \sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,\beta^{1/2+\ell(1/2-1/\beta)}\,.
\end{aligned}$$

The above series is summable whenever $\beta \in (1,2)$. For instance we may choose $\beta = 3/2$, for which

$$\sum_{\ell=0}^{+\infty}\Big(\frac{3}{2}\Big)^{1/2+\ell(1/2-2/3)} = \sqrt{\frac{3}{2}}\sum_{\ell=0}^{+\infty}\alpha^\ell = \frac{1}{1-\alpha}\sqrt{\frac{3}{2}} \leqslant 19 \qquad \text{where} \quad \alpha = \Big(\frac{3}{2}\Big)^{(1/2-2/3)} \in (0,1)$$

Therefore we have shown that

$$\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_t}(t)\big)^+ \mathbb{1}_{\{N_{a^\star}(t)<\tau_t/K\}}\Big] \leqslant 19\sqrt{\frac{\pi}{8}}\sqrt{\frac{K}{\tau_t}}\,. \tag{2.44}$$

Combining this bound with (2.43) and summing over t, we proved that the first sum in (2.42) is bounded as

$$\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star - U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big)^+\Big] \leqslant 20\sqrt{\frac{\pi}{8}}\sum_{t=K}^{T-1}\sqrt{\frac{K}{\tau_t}} \tag{2.45}$$

Remark 4. *The proof technique reveals that the bound* (2.45) *obtained in this step of the proof actually holds even if the arms are pulled according to a strategy that is not a generic MOSS strategy. This is because we never used which specific arms A^{GM}_t were pulled: we only distinguished according to how many times $a^\star$ was pulled and resorted to optional skipping.*

Control of the right deviations of all arms, that is, of the second sum in (2.42). As $(x+y)^+ \leqslant x^+ + y^+$ for all real numbers x, y, and as φ vanishes on $(0,1]$, we have, for all a and $n \geqslant 1$,

$$\begin{aligned}
\big(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\big)^+ &\leqslant \big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+ + \sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)} \\
&= \big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+ + \begin{cases} 0 & \text{if } n \geqslant T/K \\ \sqrt{\dfrac{1}{2n}\,\varphi\Big(\dfrac{T}{Kn}\Big)} & \text{if } n < T/K\,. \end{cases}
\end{aligned}$$

Therefore, for each arm a,

$$\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n} - \mu_a - \sqrt{K/T}\big)^+\Big] \leqslant \sum_{n=1}^{T}\mathbb{E}\Big[\big(\widehat{\mu}_{a,n} - \mu_a - \sqrt{K/T}\big)^+\Big] + \sum_{n=1}^{\lfloor T/K\rfloor}\sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}\,. \tag{2.46}$$

We are left with two pieces to deal with separately. For the first sum in (2.46), we exploit the integrated version of Hoeffding's inequality (Corollary 2.1),

$$\sum_{n=1}^{T}\mathbb{E}\Big[\big(\widehat{\mu}_{a,n}-\mu_a-\sqrt{K/T}\big)^+\Big]\leqslant\sqrt{\frac{\pi}{8}}\sum_{n=1}^{T}\sqrt{\frac{1}{n}}\mathrm{e}^{-2n\left(\sqrt{K/T}\right)^2}\leqslant\sqrt{\frac{\pi}{8}}\int_0^T\sqrt{\frac{1}{x}}\,\mathrm{e}^{-2xK/T}\mathrm{d}x$$
$$=\sqrt{\frac{\pi}{8}}\sqrt{\frac{T}{2K}}\int_0^{+\infty}\frac{\mathrm{e}^{-u}}{\sqrt{u}}\mathrm{d}u=\frac{\pi}{4}\sqrt{\frac{T}{K}}\,, \tag{2.47}$$

where we used the equalities $\displaystyle\int_0^{+\infty}\big(\mathrm{e}^{-u}/\sqrt{u}\big)\mathrm{d}u=2\int_0^{+\infty}\mathrm{e}^{-v^2}\mathrm{d}v=\sqrt{\pi}$.

For the second sum in (2.46), we also resort to a sum–integral comparison (which exploits the fact that φ is increasing) and perform the change of variable $u=T/(Kx)$:

$$\sum_{n=1}^{\lfloor T/K\rfloor}\sqrt{\frac{1}{2n}\,\varphi\Big(\frac{T}{Kn}\Big)}\leqslant\int_0^{T/K}\sqrt{\frac{1}{2x}\,\varphi\Big(\frac{T}{Kx}\Big)}\mathrm{d}x=\sqrt{\frac{T}{2K}}\int_1^{+\infty}u^{-3/2}\sqrt{\varphi(u)}\mathrm{d}u\,.$$

Conclusion. Getting back to (2.41) and (2.42) and collecting all the bounds above, we showed the desired bounds,

$$R_T\leqslant(K-1)\quad+\underbrace{\sum_{t=K+1}^{T}\mathbb{E}\Big[\big(\mu^\star-U^{\mathrm{GM}}_{a^\star,\tau_{t-1}}(t-1)\big)^+\Big]}_{\leqslant}\quad+\underbrace{\sum_{t=K+1}^{T}\mathbb{E}\Big[U^{\mathrm{GM}}_{A_t^{\mathrm{GM}},\tau_{t-1}}(t-1)-\mu_{A_t^{\mathrm{GM}}}\Big]}_{\leqslant}$$
$$\leqslant(K-1)\quad+\quad 20\sqrt{\frac{\pi}{8}}\sum_{t=K}^{T-1}\sqrt{\frac{K}{\tau_t}}\qquad\qquad+\underbrace{\sqrt{KT}+\sum_{a=1}^{K}\sum_{n=1}^{T}\mathbb{E}\Big[\big(U^{\mathrm{GM}}_{a,T,n}-\mu_a-\sqrt{K/T}\big)^+\Big]}$$
$$\leqslant(K-1)\quad+\quad\underbrace{20\sqrt{\frac{\pi}{8}}}_{\leqslant 12.6}\sum_{t=K}^{T-1}\sqrt{\frac{K}{\tau_t}}\qquad\qquad+\sqrt{KT}\left(1+\frac{\pi}{4}+\frac{1}{\sqrt{2}}\int_1^{+\infty}u^{-3/2}\sqrt{\varphi(u)}\mathrm{d}u\right).$$

□

2.B. Proofs of the regularity and deviation/concentration results on $\mathcal{K}_{\inf}$

We provide here the proofs of all claims made in Section 2.4.4 about the $\mathcal{K}_{\inf}$ function. These proofs are all standard but we occasionally provide simpler or more direct arguments (or slightly refined bounds).

2.B.1. Proof of the regularity lemma (Lemma 2.1)

The proof below is a variation on the proofs that can be found in Honda and Takemura [2015] or earlier references of the same authors.

Proof. To prove (2.19) we lower bound $\mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. To that end, given the definition (2.2), we lower bound $\mathrm{KL}(\nu, \nu')$ for any fixed probability distribution $\nu' \in \mathcal{P}[0,1]$ such that

$$\mathrm{E}(\nu') > \mu - \varepsilon \quad \text{and} \quad \nu' \gg \nu \,.$$

Since ν' has a countable number of atoms, one can pick a real number $x > \mu$, arbitrary close to 1, such that $\delta_x \perp \nu'$ (such that the two probability measures δ_x and ν' are singular), where δ_x is the Dirac distribution at x. We define

$$\nu'_\alpha = (1-\alpha)\nu' + \alpha\delta_x \qquad \text{where} \qquad \alpha = \frac{\varepsilon}{\varepsilon + (x-\mu)} \in (0,1)\,.$$

The expectation of ν'_α satisfies

$$\mathrm{E}(\nu'_\alpha) = (1-\alpha)\,\mathrm{E}(\nu') + \alpha x > (1-\alpha)(\mu-\varepsilon) + \alpha x = \frac{(x-\mu)(\mu-\varepsilon)}{\varepsilon + (x-\mu)} + \frac{\varepsilon x}{\varepsilon + (x-\mu)} = \mu\,.$$

Since $\alpha \in (0,1)$, we have $\nu'_\alpha \gg \nu'$; therefore, $\nu'_\alpha \gg \nu' \gg \nu$ and $\delta_x \perp \nu'$, which imply the following equalities involving densities (Radon-Nikodym derivatives):

$$\frac{\mathrm{d}\nu'}{\mathrm{d}\nu'_\alpha} = \frac{1}{1-\alpha} \qquad \text{thus} \qquad \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha} = \frac{\mathrm{d}\nu'}{\mathrm{d}\nu'_\alpha}\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} = \frac{1}{1-\alpha}\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\,. \tag{2.48}$$

This allows to compute explicitly the following Kullback-Leibler divergence:

$$\mathrm{KL}(\nu, \nu'_\alpha) = \int_{[0,1]} \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\right)\mathrm{d}\nu = \mathrm{KL}(\nu,\nu') + \ln\frac{1}{1-\alpha}\,.$$

Since $\mathrm{E}(\nu'_\alpha) > \mu$ and by the definition of $\mathcal{K}_{\inf}$ as an infimum,

$$\mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathrm{KL}(\nu,\nu'_\alpha) = \mathrm{KL}(\nu,\nu') + \ln\frac{1}{1-\alpha}\,.$$

Letting x go to 1, which implies that α goes to $\varepsilon/(1-\mu+\varepsilon)$, yields

$$\mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathrm{KL}(\nu,\nu') + \ln\frac{1-\mu+\varepsilon}{1-\mu} = \mathrm{KL}(\nu,\nu') + \ln\left(1 + \frac{\varepsilon}{1-\mu}\right) \leqslant \mathrm{KL}(\nu,\nu') + \frac{\varepsilon}{1-\mu}$$

where we also used $\ln(1+u) \leqslant u$ for all $u > -1$. Finally, by taking the infimum in the right-most equation above over all probability distributions ν' such that $\mathrm{E}(\nu') > \mu - \varepsilon$ and $\nu' \gg \nu$, we obtain the desired inequality

$$\mathcal{K}_{\inf}(\nu,\mu) \leqslant \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) + \frac{\varepsilon}{1-\mu}\,.$$

To prove the second part (2.20) of Lemma 2.1, we follow a similar path as above. We lower bound $\mathrm{KL}(\nu, \nu')$ for any fixed probability distribution $\nu' \in \mathcal{P}[0,1]$ such that

$$\mathrm{E}(\nu') > \mu \qquad \text{and} \qquad \nu' \gg \nu\,.$$

To that end, we introduce

$$\nu'_\alpha = (1-\alpha)\nu' + \alpha\nu \qquad \text{for} \qquad \alpha = \frac{\varepsilon}{\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big)} \in (0,1)$$

where $\alpha \in (0,1)$ since $\mathrm{E}(\nu) \leqslant \mu - \varepsilon$ by assumption and $\mathrm{E}(\nu') > \mu$. These two inequalities also indicate that

$$\mathrm{E}(\nu') - \mathrm{E}(\nu) > \varepsilon \qquad \text{thus} \qquad \mathrm{E}(\nu'_\alpha) = \mathrm{E}(\nu') - \alpha\big(\mathrm{E}(\nu') - \mathrm{E}(\nu)\big) > \mu - \varepsilon \tag{2.49}$$

so that $\mathrm{KL}(\nu, \nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$. Now, thanks to the absolute continuities $\nu' \gg \nu'_\alpha \gg \nu$, we have

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\frac{\mathrm{d}\nu'_\alpha}{\mathrm{d}\nu'} = \frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big((1-\alpha) + \alpha\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\,.$$

Therefore, by Fubini's theorem, the Kullback-Leibler divergence between ν and ν' equals

$$\begin{aligned}
\mathrm{KL}(\nu, \nu') = \int_{[0,1]} \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\mathrm{d}\nu &= \int_{[0,1]} \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\mathrm{d}\nu + \int_{[0,1]} \ln\Big((1-\alpha) + \alpha\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\mathrm{d}\nu \\
&\geqslant \int_{[0,1]} \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'_\alpha}\Big)\mathrm{d}\nu + \alpha\int_{[0,1]} \ln\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\Big)\mathrm{d}\nu \\
&= \mathrm{KL}(\nu, \nu'_\alpha) + \alpha\,\mathrm{KL}(\nu, \nu')
\end{aligned}$$

where we use the concavity of logarithm for the inequality. By Pinsker's inequality together with the data-processing inequality for Kullback-Leibler divergences (see, e.g., Garivier et al., 2019, Lemma 1),

$$\mathrm{KL}(\nu, \nu') \geqslant \mathrm{KL}\Big(\mathrm{Ber}\big(\mathrm{E}(\nu)\big), \mathrm{Ber}\big(\mathrm{E}(\nu')\big)\Big) \geqslant 2\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2\,.$$

Substituting this inequality above, we proved so far

$$\mathrm{KL}(\nu, \nu') \geqslant \mathrm{KL}(\nu, \nu'_\alpha) + \alpha\,\mathrm{KL}(\nu, \nu') \geqslant \mathrm{KL}(\nu, \nu'_\alpha) + 2\alpha\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)^2 = \mathrm{KL}(\nu, \nu'_\alpha) + 2\varepsilon\big(\mathrm{E}(\nu) - \mathrm{E}(\nu')\big)$$

where we used the definition of α for the last inequality. By applying the bound (2.49) and its consequence $\mathrm{KL}(\nu, \nu'_\alpha) \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon)$, we finally get

$$\mathrm{KL}(\nu, \nu') \geqslant \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) + 2\varepsilon^2\,.$$

The proof of (2.20) is concluded by taking the infimum in the left-hand side over the probability distributions ν' such that $\mathrm{E}(\nu') > \mu$ (and $\nu' \gg \nu$). □

2.B.2. A useful tool: a variational formula for $\mathcal{K}_{\inf}$ (statement)

The variational formula below appears in Honda and Takemura [2015] as Theorem 2 (and Lemma 6) and is an essential tool for deriving the deviation and concentration results for the $\mathcal{K}_{\inf}$. We state it here (and re-derive it in a direct way in Appendix 2.D) for the sake of completeness.

Lemma 2.3 (variational formula for $\mathcal{K}_{\inf}$). *For all $\nu \in \mathcal{P}[0,1]$ and all $0 < \mu < 1$,*

$$\mathcal{K}_{\inf}(\nu, \mu) = \max_{0 \leqslant \lambda \leqslant 1} \mathbb{E}\left[\ln\left(1 - \lambda \frac{X - \mu}{1 - \mu}\right)\right] \qquad \text{where } X \sim \nu\,. \tag{2.50}$$

Moreover, if we denote by $\lambda^\star$ the value at which the above maximum is reached, then

$$\mathbb{E}\left[\frac{1}{1 - \lambda^\star (X - \mu)/(1 - \mu)}\right] \leqslant 1\,. \tag{2.51}$$

2.B.3. Proof of the deviation result (Proposition 2.4)

The following proof is almost exactly the same as that of Cappé et al. [2013, Lemma 6], except that we correct a small mistake in the constant.

Proof. We first upper bound $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)$: as indicated by the variational formula of Lemma 2.3, it is a maximum of random variables indexed by $[0,1]$. We provide an upper bound that is a finite maximum. To that end, we fix a real number $\gamma \in (0,1)$, to be determined by the analysis, and let S_γ be the set

$$S_\gamma = \left\{\frac{1}{2} - \left\lfloor \frac{1}{2\gamma} \right\rfloor \gamma, \ldots, \frac{1}{2} - \gamma,\ \frac{1}{2},\ \frac{1}{2} + \gamma, \ldots, \frac{1}{2} + \left\lfloor \frac{1}{2\gamma} \right\rfloor \gamma\right\}.$$

The cardinality of this set S_γ is bounded by $1 + 1/\gamma$. Lemma 2.4 below (together with the consequence mentioned after its statement) indicates that for all $\lambda \in [0,1]$, there exists a $\lambda' \in S_\gamma$ such that for all $x \in [0,1]$,

$$\ln\left(1 - \lambda \frac{x - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \leqslant 2\gamma + \ln\left(1 - \lambda' \frac{x - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right). \tag{2.52}$$

(The small correction with respect to the original proof is the 2γ factor in the inequality above, instead of the claimed γ term therein; this is due to the constraint $\lambda \leqslant \lambda' \leqslant 1/2$ or $1/2 \leqslant \lambda' \leqslant \lambda$ in the statement of Lemma 2.4.) Now, a combination of the variational formula of Lemma 2.3 and of the inequality (2.52) yields a finite maximum as an upper bound on $\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big)$:

$$\begin{aligned}
\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) &= \max_{0 \leqslant \lambda \leqslant 1} \frac{1}{n} \sum_{k=1}^{n} \ln\left(1 - \lambda \frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \\
&\leqslant 2\gamma + \max_{\lambda' \in S_\gamma} \frac{1}{n} \sum_{k=1}^{n} \ln\left(1 - \lambda' \frac{X_k - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right).
\end{aligned}$$

In the second part of the proof, we control the deviations of the upper bound obtained. A union bound yields

$$\begin{aligned}
\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n, \mathrm{E}(\nu)\big) \geqslant u\Big] &\leqslant \mathbb{P}\left[\max_{\lambda' \in S_\gamma} \frac{1}{n} \sum_{k=1}^{n} \ln\left(1 - \lambda' \frac{X_k - \mathrm{E}(\mu)}{1 - \mathrm{E}(\mu)}\right) \geqslant u - 2\gamma\right] \\
&\leqslant \sum_{\lambda' \in S_\gamma} \mathbb{P}\left[\frac{1}{n} \sum_{k=1}^{n} \ln\left(1 - \lambda' \frac{X_k - \mathrm{E}(\nu)}{1 - \mathrm{E}(\nu)}\right) \geqslant u - 2\gamma\right].
\end{aligned} \tag{2.53}$$

By the Markov–Chernov inequality, for all $\lambda' \in [0,1]$, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n}\ln\left(1-\lambda'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right)\geqslant u-2\gamma\right] \leqslant \mathrm{e}^{-n(u-2\gamma)}\,\mathbb{E}\left[\prod_{k=1}^{n}\left(1-\lambda'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right)\right]$$
$$= \mathrm{e}^{-n(u-2\gamma)}\prod_{k=1}^{n}\underbrace{\mathbb{E}\left[1-\lambda'\frac{X_k-\mathrm{E}(\nu)}{1-\mathrm{E}(\nu)}\right]}_{=1} = \mathrm{e}^{-n(u-2\gamma)}$$

where we used the independence of the X_k. Substituting in (2.53) and using the bound $1+1/\gamma$ on the cardinality of S_γ, we get

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_n,\mathrm{E}(\nu)\big)\geqslant u\Big] \leqslant \sum_{\lambda'\in S_\gamma}\mathrm{e}^{-n(u-2\gamma)} \leqslant (1+1/\gamma)\,\mathrm{e}^{-n(u-2\gamma)}\,.$$

Taking $\gamma = 1/(2n)$ concludes the proof. □

The proof above relies on the following lemma, which is extracted from Cappé et al. [2013, Lemma 7]. Its elementary proof (not copied here) consists in bounding of derivative of $\lambda \mapsto \ln(1-\lambda c)$ and using a convexity argument.

Lemma 2.4. *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 $c \leqslant 1$,*

$$\ln(1-\lambda c) - \ln(1-\lambda' c) \leqslant 2|\lambda-\lambda'|\,.$$

A consequence not drawn by Cappé et al. [2013] is that the lemma above actually also holds for $\lambda = 1$ and $\lambda' \in [0,1)$. Indeed, by continuity and by letting $\lambda \to 1$, we get from this lemma that for all $\lambda' \in [1/2,1)$ and for all real numbers $c < 1$,

$$\ln(1-c) - \ln(1-\lambda' c) \leqslant 2(1-\lambda')\,.$$

The above inequality is also valid for $c = 1$ as the left-hand side equals $-\infty$.

2.B.4. Proof of the concentration result (Proposition 2.5)

We recall that Proposition 2.5—and actually most of its proof below—are similar in spirit to Honda and Takemura [2015, Proposition 11]. However, they are tailored to our needs. The key ingredients in the proof will be the variational formula (2.50)—again—and Lemma 2.5 below. This lemma is a concentration result for random variables that are essentially bounded from one side only; it holds for possibly negative u (there is no lower bound on the u that can be considered).

Lemma 2.5. *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables such that there exist a, $b \geqslant 0$ with*

$$Z_1 \leqslant a \quad \text{a.s.} \qquad \text{and} \qquad \mathbb{E}\big[\mathrm{e}^{-Z_1}\big] \leqslant b\,.$$

Denote $\gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$. Then Z_1 in integrable and for all real numbers $u \in \big(-\infty, \mathbb{E}[Z_1]\big)$,

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nu\right] \leqslant \begin{cases} \exp(-n\gamma/8) & \text{if } u \leqslant \mathbb{E}[Z_1]-\gamma/2 \\ \exp\Big(-n\big(\mathbb{E}[Z_1]-u\big)^2/(2\gamma)\Big) & \text{if } u > \mathbb{E}[Z_1]-\gamma/2 \end{cases}.$$

Proof of Proposition 2.5 based on Lemma 2.5

We apply Lemma 2.3. We denote by $\lambda^\star \in [0,1]$ a real number achieving the maximum in the variational formula (2.50) for $\mathcal{K}_{\inf}(\nu,\mu)$. We then introduce the random variable

$$Z = \ln\left(1 - \lambda^\star \frac{X-\mu}{1-\mu}\right) \qquad \text{where} \qquad X \sim \nu$$

and i.i.d. copies $Z_1, \dots, Z_n$ of Z. Then, $\mathcal{K}_{\inf}(\nu,\mu) = \mathbb{E}\big[Z\big]$ and by the variational formula (2.50) again,

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_n,\mu\big) \geqslant \frac{1}{n}\sum_{i=1}^n Z_i\,, \qquad \text{therefore,} \qquad \mathbb{P}\big[\mathcal{K}_{\inf}(\widehat{\nu}_n,\mu) \leqslant x\big] \leqslant \mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nx\right]$$

for all real numbers x. Now, $X \geqslant 0$ and $\lambda^\star \leqslant 1$, thus

$$Z \leqslant \ln\left(1 + \lambda^\star \frac{\mu}{1-\mu}\right) \leqslant \ln\left(\frac{1}{1-\mu}\right) \stackrel{\text{def}}{=} a\,.$$

On the other hand,

$$\mathbb{E}\big[\mathrm{e}^{-Z}\big] = \mathbb{E}\left[\frac{1}{1-\lambda^\star(X-\mu)/(1-\mu)}\right] \stackrel{\text{def}}{=} b$$

where $b \leqslant 1$ follows from (2.51). This proves Proposition 2.5 via Lemma 2.5, except for the inequality $\mathrm{e}^{-n\gamma/8} \leqslant \mathrm{e}^{-n/4}$ claimed therein. The latter is a consequence of $\gamma \geqslant 2$, as γ is an increasing function of $\mu > 0$,

$$\gamma = \frac{1}{\sqrt{1-\mu}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu}\right)\right) > 16\mathrm{e}^{-2} > 2$$

Remark 5. *In the proof of Theorem 2.3 provided in Section 2.C we will not use Proposition 2.5 as stated but a stronger result, namely that the bound of Proposition 2.5 actually holds for the larger quantity*

$$\mathbb{P}\left[\sum_{i=1}^n Z_i \leqslant nx\right]$$

as is clear from the proof above.

Proof of Lemma 2.5

This lemma is a direct application of the Crámer–Chernov method. We introduce the log-moment generating function Λ of Z_1:

$$\Lambda : x \longmapsto \ln \mathbb{E}\big[\mathrm{e}^{xZ_1}\big]\,.$$

Lemma 2.6. *The log-moment generating function Λ is well-defined at least on the interval $[-1,1]$ and twice differentiable at least on $(-1,1)$, with $\Lambda'(0) = \mathbb{E}[Z_1]$ and $\Lambda''(x) \leqslant \gamma$ for $x \in [-1/2,\, 0]$, where $\gamma = \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big)$ denotes the same constant as in Lemma 2.5.*

Based on this lemma (proved below), we may resort to a Taylor expansion with a Lagrange remainder and get the bound:

$$\forall\, x \in [-1/2,\, 0], \qquad \Lambda(x) \leqslant \Lambda(0) + x\,\Lambda'(0) + \frac{x^2}{2}\sup_{y\in[-1/2,\,0]}\Lambda''(y) \leqslant x\,\mathbb{E}[Z_1] + \frac{\gamma}{2}x^2\,.$$

Therefore, by the Crámer–Chernov method, for all $x \in [-1/2,\, 0]$, the probability of interest is bounded by

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nu\right] = \mathbb{P}\left[\prod_{i=1}^{n} \mathrm{e}^{xZ_i} \geqslant \mathrm{e}^{nux}\right] \leqslant \mathrm{e}^{-nux}\left(\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]\right)^n = \exp\Big(-n\big(ux - \Lambda(x)\big)\Big)$$
$$\leqslant \exp\Big(n\Big(x^2\,\gamma/2 - x\,\big(u - \mathbb{E}[Z_1]\big)\Big)\Big)\,. \tag{2.54}$$

That is,

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i \leqslant nu\right] \leqslant \exp\Big(n \min_{x\in[1/2,\,0]} P(x)\Big)$$

where we introduced the second-order polynomial function

$$P(x) = x^2\,\gamma/2 - x\,\big(u - \mathbb{E}[Z_1]\big) = \frac{\gamma x}{2}\left(x - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right).$$

The claimed bound is obtained by minimizing P over $[-1/2,\, 0]$ depending on whether $u > \mathbb{E}[Z_1] - \gamma/2$ or $u \leqslant \mathbb{E}[Z_1] - \gamma/2$, which we do now.

We recall that by assumption, $u < \mathbb{E}[Z_1]$. We note that P is a second-order polynomial function with positive leading coefficient and roots 0 and $2\big(u - \mathbb{E}[Z_1]\big)/\gamma < 0$. Its 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 P is to be minimized over $[-1/2,\, 0]$ only. In the case where $u > \mathbb{E}[Z_1] - \gamma/2$, the midpoint $x^\star$ belongs to the interval of interest and

$$\min_{[-1/2,0]} P = \frac{\gamma x^\star}{2}\left(x^\star - 2\frac{u - \mathbb{E}[Z_1]}{\gamma}\right) = -\frac{\big(u - \mathbb{E}[Z_1]\big)^2}{2\gamma}\,.$$

Otherwise, $u - \mathbb{E}[Z_1] \leqslant -\gamma/2$ and the midpoint $x^\star$ is to the left of $-1/2$. Therefore, P is increasing on $[-1/2, 0]$, so that its minimum on this interval is achieved at $-1/2$, that is,

$$\min_{[-1/2,0]} P = P(-1/2) = \frac{\gamma}{8} + \frac{1}{2}\big(u - \mathbb{E}[Z_1]\big) \leqslant \frac{\gamma}{8} - \frac{\gamma}{4} = -\frac{\gamma}{8}\,.$$

This concludes the proof of Lemma 2.5. We end this section by proving Lemma 2.6, which stated some properties of the Λ function.

Proof of Lemma 2.6. 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,

$$\mathbb{E}\big[|Z_1|\big] \leqslant \ln \mathbb{E}\Big[\mathrm{e}^{|Z_1|}\Big] \leqslant \ln\Big(\mathbb{E}\big[\mathrm{e}^{-Z_1}\big] + \mathbb{E}\big[\mathrm{e}^{Z_1}\big]\Big) < +\infty\,.$$

First, 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.

Second, 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

$$\psi'(x) = \mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big] \qquad \text{and} \qquad \psi''(x) = \mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]\,.$$

Therefore, so is $\Lambda = \ln\psi$, with derivatives

$$\Lambda'(x) = \frac{\psi'(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]} \quad \text{and} \quad \Lambda''(x) = \frac{\psi''(x)\,\psi(x) - \big(\psi'(x)\big)^2}{\psi(x)^2} \leqslant \frac{\psi''(x)}{\psi(x)} = \frac{\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]}\,.$$

In particular, $\Lambda'(0) = \mathbb{E}[Z_1]$.

Finally, for the bound on $\Lambda''(x)$, we note first that $Z_1 \leqslant a$ (with $a \geqslant 0$) and $x \in [-1/2,\,0]$ entail that $\mathrm{e}^{xZ_1} \geqslant \mathrm{e}^{xa} \geqslant 1/\sqrt{\mathrm{e}^a}$. Second, $\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big] \leqslant 16\,\mathrm{e}^{-2}b + a^2$ follows from replacing z by Z_1 and taking expectations in the inequality (proved below)

$$\forall\, x \in [-1/2,\,0],\ z \in (-\infty, a], \qquad z^2\,\mathrm{e}^{xz} \leqslant 16\,\mathrm{e}^{-2}\mathrm{e}^{-z} + a^2\,. \tag{2.55}$$

Collecting all elements together, we proved

$$\Lambda''(x) \leqslant \frac{\mathbb{E}\big[Z_1^2\,\mathrm{e}^{xZ_1}\big]}{\mathbb{E}\big[\mathrm{e}^{xZ_1}\big]} \leqslant \sqrt{\mathrm{e}^a}\big(16\,\mathrm{e}^{-2}b + a^2\big) = \gamma\,.$$

To see why (2.55) holds, note that in the case $z \geqslant 0$, since $x \leqslant 0$ we have $z^2\,\mathrm{e}^{xz} \leqslant z^2 \leqslant a^2$. In the case $z \leqslant 0$, we have (by function study) $z^2 \leqslant 16\mathrm{e}^{-2-z/2}$, so that $z^2\,\mathrm{e}^{xz} \leqslant 16\mathrm{e}^{-2}\,\mathrm{e}^{(x-1/2)z} \leqslant 16\mathrm{e}^{-2}\mathrm{e}^{-z}$ where we used $x \geqslant -1/2$ for the final inequality. □

2.C. Proof of Theorem 2.3 (with the $-\ln\ln T$ term in the regret bound)

We incorporate two refinements to the proof of Theorem 2.2 in Section 2.6.2 to obtain Theorem 2.3 with this improved $-\ln\ln T$ term. First, the left deviations of the index are controlled with an additional cut on the value of $U_a(t)$ *before* using the bound $U_a(t) \geqslant U_{a^\star}(t)$ that holds when $A_{t+1} = a$. This improves the dependency on the parameter δ used in the proof; as a consequence, $\delta = T^{-1/8}$ will be set instead of $\delta = (\ln T)^{-1/3}$, which will improve the order of magnitude of second-order terms. Second, to sharpen the bound on the quantity (2.60), which contains the main logarithmic term, we use a trick introduced in the analysis of the IMED policy by Honda and Takemura [2015, Theorem 5]. Their idea was to deal with the deviations in a more careful way and relate the sum (2.60) to the behaviour of a biased random walk. Doing so, we obtain a bound of the form $\kappa\, W(cT)$, where W is Lambert's function, instead of the bound of the form $\kappa \ln(cT)$ stated in Theorem 2.2.

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 $w > 0$. It is an increasing function satisfying (see, e.g., Hoorfar and Hassani, 2008, Corollary 2.4)

$$\forall\, x > \mathrm{e}, \qquad \ln x - \ln\ln x \leqslant W(x) \leqslant \ln x - \ln\ln x + \ln\left(1 + \mathrm{e}^{-1}\right). \tag{2.56}$$

In particular, $W(x) = \ln x - \ln\ln x + \mathcal{O}(1)$ as $x \to +\infty$.

What we will exactly prove below is the following. We recall that we assume here $\mu^\star \in (0,1)$. Given $T \geqslant K/(1-\mu^\star)$, 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, and for all $\delta > 0$ satisfying

$$\delta < \min\left\{\mu^\star,\ \frac{\Delta_a}{2},\ \frac{1-\mu^\star}{2}\,\mathcal{K}_{\inf}(\nu_a, \mu^\star)\right\}$$

we have

$$\begin{aligned}
\mathbb{E}[N_a(T)] \leqslant\ & 1 \\
& + \frac{5\mathrm{e}K}{\left(1-\mathrm{e}^{-\Delta_a^2/2}\right)^3} + T\mathrm{e}^{-\Delta_a^2 T/(2K)} \\
& + \frac{K/T}{1-\mathrm{e}^{-\Delta_a^2/8}} \\
& + \left\lceil \frac{8}{\Delta_a^2}\,\ln\left(\frac{T}{K}\right)\right\rceil \left(\frac{5\mathrm{e}K/T}{(1-\mathrm{e}^{-2\delta^2})^3} + \mathrm{e}^{-2\delta^2 T/K}\right) \\
& + \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)} \left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \\
& \qquad\qquad + 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \\
& + \frac{1}{1-\mathrm{e}^{-\Delta_a^2/8}}\,.
\end{aligned} \tag{2.57}$$

We write the bound in this way to match the decomposition of $\mathbb{E}[N_a(T)]$ appearing in the proof (see page 73). For a choice $\delta \to 0$ $T \to +\infty$, the previous bound is of the form

$$\mathbb{E}[N_a(T)] \leqslant \frac{W\big(c_{\mu^\star} T\big)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \delta/(1-\mu^\star)} + \mathcal{O}_T\left(\frac{\ln T}{\delta^6 T}\right) + \mathcal{O}_T\big((\ln T)\,\mathrm{e}^{-2\delta^2 T/K}\big) + \mathcal{O}_T(1)$$

where $c_{\mu^\star} = \ln\bigl(1/(1-\mu^\star)\bigr)/K$. Based on the first-order approximation $1/(1-\varepsilon) = 1+\varepsilon+\mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$ and on the inequalities (2.56), we get

$$\mathbb{E}[N_a(T)] \leqslant \frac{\ln T - \ln\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\bigl(1+\mathcal{O}_T(\delta)\bigr) + \mathcal{O}_T\left(\frac{\ln T}{\delta^6 T}\right) + \mathcal{O}_T\bigl((\ln T)\,\mathrm{e}^{-2\delta^2 T/K}\bigr) + \mathcal{O}_T(1)\,.$$

The choice $\delta = T^{-1/8}$ leads to the bound stated in Theorem 2.3, namely,

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

We now prove the closed-form bound (2.57).

Proof. As in the proof of Theorem 2.2, given $\delta > 0$ sufficiently small, we decompose $\mathbb{E}\bigl[N_a(T)\bigr]$. However, this time we refine the decomposition quite a bit. Instead of simply distinguishing whether $U_a(t)$ is greater or smaller than $\mu^\star - \delta$, we add a cutting point at $(\mu^\star + \mu_a)/2$. 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 $N_a(t) \leqslant n_0 - 1$ when $U_a(t) < \mu^\star - \delta$, while we keep the integer threshold $f(T,K)$ in the case $U_a(t) \geqslant \mu^\star - \delta$. More precisely,

$$\begin{aligned}
\bigl\{U_a(t) < \mu^\star - \delta\bigr\} \cup \bigl\{U_a(t) \geqslant \mu^\star - \delta\bigr\} = \quad & \bigl\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \geqslant n_0\bigr\} \\
& \cup \bigl\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \leqslant n_0 - 1\bigr\} \\
& \cup \bigl\{U_a(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \leqslant f(T,K)\bigr\} \\
& \cup \bigl\{U_a(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \geqslant f(T,K)+1\bigr\} \\
\subseteq \quad & \bigl\{U_a(t) < (\mu^\star + \mu_a)/2 \text{ and } N_a(t) \geqslant n_0\bigr\} \\
& \cup \bigl\{(\mu^\star + \mu_a)/2 \leqslant U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \geqslant n_0\bigr\} \\
& \cup \bigl\{U_a(t) < \mu^\star - \delta \text{ and } N_a(t) \leqslant n_0 - 1\bigr\} \\
& \cup \bigl\{U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \leqslant f(T,K)\bigr\} \\
& \cup \bigl\{U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \text{ and } N_a(t) \geqslant f(T,K)+1\bigr\}
\end{aligned}$$

where, to get the inclusion, we further cut the first event into two events and we used the definition of the index $U_a(t)$ to replace it by $U_a^{\mathrm{KL}}(t)$ or $U_a^{\mathrm{M}}(t)$ in the last two events.

Hence, by intersecting this partition of the space with the event $\{A_{t+1} = a\}$ and by slightly simplifying the first and second events of the partition:

$$\begin{aligned}
\{A_{t+1} = a\} \subseteq \quad & \bigl\{U_a(t) < (\mu^\star + \mu_a)/2 \text{ and } A_{t+1} = a\bigr\} \\
& \cup \bigl\{U_a(t) \geqslant (\mu^\star + \mu_a)/2 \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\bigr\} \\
& \cup \bigl\{U_a(t) < \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant n_0 - 1\bigr\} \\
& \cup \bigl\{U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \leqslant f(T,K)\bigr\} \\
& \cup \bigl\{U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant f(T,K)+1\bigr\}
\end{aligned}$$

Only now do we inject the bound $U_{a^\star}(t) \leqslant U_a(t)$, valid when $A_{t+1} = a$, as well as a union

bound, to obtain our working decomposition of $\mathbb{E}[N_a(t)]$:

$$\begin{aligned}
\mathbb{E}\big[N_a(T)\big] \leqslant\ & 1 \\
& + \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}(t) < (\mu^\star + \mu_a)/2\big] && (S_1) \\
& + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a(t) \geqslant (\mu^\star + \mu_a)/2 \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \geqslant n_0\big] && (S_2) \\
& + \sum_{t=K}^{T-1} \mathbb{P}\big[U_{a^\star}(t) < \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant n_0 - 1\big] && (S_3) \\
& + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{KL}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(T,K)\big] && (S_4) \\
& + \sum_{t=K}^{T-1} \mathbb{P}\big[U_a^{\mathrm{M}}(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \geqslant f(T,K) + 1\big]\,. && (S_5)
\end{aligned}$$

We call the five sums appearing in the right-hand side S_1, S_2, S_3, S_4, S_5, respectively and now bound them separately. Most of the efforts will be dedicated to the sum S_4.

Bound on S_5

As the algorithm considered is the same as in Theorem 2.2, its analysis is still valid. Fortunately, the S_5 term was already covered in (2.38): provided that $\delta < \Delta_a/4$,

$$S_5 \leqslant \frac{1}{1 - \mathrm{e}^{-\Delta_a^2/8}}\,.$$

Bound on S_2

Let

$$n_0 = \left\lceil \frac{8}{\Delta_a^2} \ln\!\left(\frac{T}{K}\right) \right\rceil. \tag{2.58}$$

By Pinsker's inequality (2.8), by definition of the MOSS index, and by our choice of n_0, we have, when $N_a(t) \geqslant n_0$,

$$U_a(t) \leqslant U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{2N_a(t)} \ln_+\!\left(\frac{T}{KN_a(t)}\right)} \leqslant \widehat{\mu}_a(t) + \underbrace{\sqrt{\frac{1}{2n_0} \ln_+\!\left(\frac{T}{Kn_0}\right)}}_{\leqslant \Delta_a/4}. \tag{2.59}$$

In particular, we get the inclusion

$$\begin{aligned}
\big\{U_a(t) \geqslant (\mu^\star + \mu_a)/2 \ \text{ and } \ N_a(t) \geqslant n_0\big\} &= \big\{U_a(t) \geqslant \mu_a + \Delta_a/2 \ \text{ and } \ N_a(t) \geqslant n_0\big\} \\
&\subseteq \big\{\widehat{\mu}_a(t) \geqslant \mu_a + \Delta_a/4 \ \text{ and } \ N_a(t) \geqslant n_0\big\}\,.
\end{aligned}$$

Thus

$$S_2 \leqslant \sum_{t=K}^{T-1} \mathbb{P}\left[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \geqslant n_0\right].$$

We now proceed similarly to what we already did on page 54. By a careful application of optional skipping (see Section 2.4.1), using the fact that all the events $\{A_{t+1} = a \text{ and } N_a(t) = n\}$ are disjoint as t varies, the sum above may be bounded by

$$\sum_{t=K}^{T-1} \mathbb{P}\Big[\widehat{\mu}_a(t) \geqslant \mu_a + \frac{\Delta_a}{4} \text{ and } A_{t+1} = a \text{ and } N_a(t) \geqslant n_0\Big] \leqslant \sum_{n \geqslant n_0} \mathbb{P}\Big[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\Big]$$

By a final application of Hoeffding's inequality (Proposition 2.1, actually not using the maximal form):

$$S_2 \leqslant \sum_{n=n_0}^{T} \mathbb{P}\Big[\widehat{\mu}_{a,n} \geqslant \mu_a + \frac{\Delta_a}{4}\Big] \leqslant \sum_{n=n_0}^{T} \mathrm{e}^{-n\Delta_a^2/8} = \frac{\mathrm{e}^{-n_0\Delta_a^2/8}}{1-\mathrm{e}^{-\Delta_a^2/8}} \leqslant \frac{K/T}{1-\mathrm{e}^{-\Delta_a^2/8}}$$

where we substituted the value (2.58) of n_0.

Bounds on S_1 and S_3

For $u \in (0,1)$, we introduce the event

$$\mathcal{E}_\star(u) = \Big\{\exists\, \tau \in \{K, \dots, T-1\} :\ U_{a^\star}(\tau) < u\Big\}$$

so that

$$\{U_{a^\star}(t) < (\mu^\star + \mu_a)/2\} \subseteq \mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big) \qquad \text{and} \qquad \{U_{a^\star}(t) < \mu^\star - \delta\} \subseteq \mathcal{E}_\star\big(\mu^\star - \delta\big)\,.$$

Summing over t, and using the deterministic control

$$\sum_{t=K}^{T-1} \mathbb{1}_{\{A_{t+1}=a \text{ and } N_a(t) \leqslant n_0 - 1\}} \leqslant n_0$$

for bounding S_3, we obtain (and this is where it is handy that the $\mathcal{E}_\star$ do not depend on a particular t)

$$S_1 \leqslant T\ \mathbb{P}\Big(\mathcal{E}_\star\big((\mu^\star + \mu_a)/2\big)\Big) \qquad \text{and} \qquad S_3 \leqslant n_0\ \mathbb{P}\big(\mathcal{E}_\star(\mu^\star - \delta)\big)$$

We recall that n_0 was defined in (2.58). The lemma right below, respectively with $x = \Delta_a/2$ and $x = \delta$, yield the final bounds

$$S_1 \leqslant \frac{5\mathrm{e}K}{\big(1-\mathrm{e}^{-\Delta_a^2/2}\big)^3} + T\mathrm{e}^{-\Delta_a^2 T/(2K)}$$

and

$$S_3 \leqslant \left\lceil \frac{8}{\Delta_a^2} \ln\left(\frac{T}{K}\right)\right\rceil \left(\frac{5\mathrm{e}K/T}{(1-\mathrm{e}^{-2\delta^2})^3} + \mathrm{e}^{-2\delta^2 T/K}\right).$$

Lemma 2.7. *For all $x \in (0, \mu^\star)$,*

$$\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star - x\big)\Big) = \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ U_{a^\star}(\tau) < \mu^\star - x\Big] \leqslant \frac{\mathrm{e}K}{T}\frac{5}{(1-\mathrm{e}^{-2x^2})^3} + \mathrm{e}^{-2x^2 T/K}\,.$$

Proof. We first lower bound $U_{a^\star}(\tau)$ depending on whether $N_{a^\star}(\tau) < T/K$ or $N_{a^\star}(\tau) \geqslant T/K$. In the first case, we will simply apply Pinsker's inequality (2.8) to get $U_{a^\star}^{\mathrm{KL}}(\tau) \leqslant U_{a^\star}(\tau)$. In the second case, since $T \geqslant K/(1-\mu^\star) \geqslant K$, we have, by definition of $f(T,K)$, that $T/K \geqslant (T/K)^{1/5} \geqslant f(T,K)$ and thus, by definition of the $U_{a^\star}(\tau)$ index, $U_{a^\star}(\tau) = U_{a^\star}^{\mathrm{M}}(\tau)$. Now, the $\ln_+$

in the definition of $U^{\mathrm{M}}_{a^\star}(\tau)$ vanishes when $N_{a^\star}(\tau) \geqslant T/K$, so all in all we have $U_{a^\star}(\tau) = \widehat{\mu}_{a^\star}(\tau)$ when $N_{a^\star}(\tau) \geqslant T/K$. Therefore, by optional skipping (see Section 2.4.1),

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star - x\big)\Big) &= \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ U_{a^\star}(\tau) < \mu^\star - x\Big] \\
&= \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ U_{a^\star}(\tau) < \mu^\star - x \ \text{ and } \ N_{a^\star}(\tau) < T/K\Big] \\
&\quad + \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ U_{a^\star}(\tau) < \mu^\star - x \ \text{ and } \ N_{a^\star}(\tau) \geqslant T/K\Big] \\
&\leqslant \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ U^{\mathrm{KL}}_{a^\star}(\tau) < \mu^\star - x \ \text{ and } \ N_{a^\star}(\tau) < T/K\Big] \\
&\quad + \mathbb{P}\Big[\exists\, \tau \in \{K, \dots, T-1\} :\ \widehat{\mu}_{a^\star}(\tau) < \mu^\star - x \ \text{ and } \ N_{a^\star}(\tau) \geqslant T/K\Big] \\
&\leqslant \mathbb{P}\Big[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} :\ U^{\mathrm{KL}}_{a^\star,m} < \mu^\star - x\Big] \\
&\quad + \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} :\ \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] .
\end{aligned}$$

As in the proof of Corollary 2.4, by the definition of the $U^{\mathrm{KL}}_{a^\star,m}$ index as some supremum (together with the left-continuity of $\mathcal{K}_{\inf}$ deriving from Lemma 2.1), we finally get

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}_\star\big(\mu^\star - x\big)\Big) \leqslant \mathbb{P}\Bigg[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} :\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m} \ln\left(\frac{T}{Km}\right)\Bigg] \\
+ \mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} :\ \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] .
\end{aligned}$$

The proof is concluded by bounding each probability separately. First, again as in the proof of Corollary 2.4, we apply Corollary 2.3 (for the first inequality below) and the deviation inequality of Proposition 2.4 (for the second inequality below), to see that for all $x \in (0, \mu^\star)$ and $\varepsilon > 0$,

$$\mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \varepsilon\Big] \leqslant \mathbb{P}\Big[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star\big) > \varepsilon + 2x^2\Big] \leqslant \mathrm{e}(2n+1)\,\mathrm{e}^{-n(\varepsilon+2x^2)} .$$

Therefore, by a union bound, the above equation, and the calculations on geometric sums (2.33) and (2.34),

$$\begin{aligned}
&\mathbb{P}\Bigg[\exists\, m \in \big\{1, \dots, \lfloor T/K \rfloor\big\} :\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a^\star,m}, \mu^\star - x\big) > \frac{1}{m} \ln\left(\frac{T}{Km}\right)\Bigg] \\
&\qquad \leqslant \sum_{m=1}^{\lfloor T/K \rfloor} \mathrm{e}(2m+1)\,\frac{Km}{T}\mathrm{e}^{-2mx^2} \leqslant \frac{\mathrm{e}K}{T} \sum_{m=1}^{+\infty} m(2m+1)\,\mathrm{e}^{-2mx^2} \leqslant \frac{\mathrm{e}K}{T}\,\frac{5}{(1-\mathrm{e}^{-2x^2})^3} .
\end{aligned}$$

Second, by Hoeffding's maximal inequality (Proposition 2.1),

$$\begin{aligned}
&\mathbb{P}\Big[\exists\, m \in \big\{\lceil T/K \rceil, \dots, T\big\} :\ \widehat{\mu}_{a^\star,m} < \mu^\star - x\Big] \\
&\qquad = \mathbb{P}\left[\max_{\lceil T/K \rceil \leqslant m \leqslant T} \Big(\big(1 - \widehat{\mu}_{a^\star,m}\big) - \big(1 - \mu^\star\big)\Big) > x\right] \leqslant \mathrm{e}^{-2\,\lceil T/K \rceil\, x^2} \leqslant \mathrm{e}^{-2x^2 T/K} .
\end{aligned}$$

The proof is concluded by collecting the last two bounds. □

Bound on S_4

We begin with a now standard use of optional skipping (see Section 2.4.1), relying on the fact that the events $\{A_{t+1} = a \ \text{ and } \ N_a(t) = n\}$ are disjoint as t varies:

$$S_4 = \sum_{t=K}^{T-1} \mathbb{P}\big[U^{\mathrm{KL}}_a(t) \geqslant \mu^\star - \delta \ \text{ and } \ A_{t+1} = a \ \text{ and } \ N_a(t) \leqslant f(T,K)\big] \leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U^{\mathrm{KL}}_{a,n} \geqslant \mu^\star - \delta\big] .$$

We show in this section that

$$\sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] \leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)} \left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\, T \right) + \ln\big(2/(1-\mu^\star)\big) \right) + 5 + \frac{1}{1 - \mathrm{e}^{-\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}} \quad (2.60)$$

where, as in the statement of Proposition 2.5,

$$\gamma_\star = \frac{1}{\sqrt{1-\mu^\star}} \left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-\mu^\star} \right) \right).$$

To do so, 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 (2.60) to the behaviour of a biased random walk.

We start by rewriting the events of interest as

$$\big\{U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big\} = \left\{ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n} \ln\left(\frac{T}{Kn} \right) \right\}$$

where, as in one step of the proof of Lemma 2.7, we used the definition of $U_{a,n}^{\mathrm{KL}}$ as well as the left-continuity of $\mathcal{K}_{\inf}$. We then follow the same steps as in the proof of Proposition 2.5 (see Section 2.B.4) and link the deviations in $\mathcal{K}_{\inf}$ divergence to the ones of a random walk. The variational formula (Lemma 2.3) for $\mathcal{K}_{\inf}$ entails the existence of $\lambda_{a,\delta} \in [0,1]$ such that

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) = \mathbb{E}\left[\ln\left(1 - \lambda_{a,\delta} \frac{X_a - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)} \right) \right] \qquad \text{where} \qquad X_a \sim \nu_a\,.$$

Note that $\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) > 0$ by (2.7) given that we imposed $\delta \leqslant \Delta_a/2$. We consider i.i.d. copies $X_{a,1}, \dots, X_{a,n}$ of X and form the random variables

$$Z_{a,i} = \ln\left(1 - \lambda_{a,\delta} \frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)} \right).$$

By the variational formula (Lemma 2.3) again, applied this time to $\mathcal{K}_{\inf}(\widehat{\nu}_{a,n}, \mu^\star - \delta)$, we see

$$\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \geqslant \frac{1}{n} \sum_{i=1}^{n} Z_{a,i}$$

which entails, for each $n \geqslant 1$,

$$\left\{ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n} \ln\left(\frac{T}{Kn} \right) \right\} \subseteq \left\{ \sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn} \right) \right\}. \quad (2.61)$$

Collecting all previous bounds and inclusions, we proved that the sum of interest (2.60) is bounded by

$$\begin{aligned} S_4 &\leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\big[U_{a,n}^{\mathrm{KL}} \geqslant \mu^\star - \delta\big] = \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\mathcal{K}_{\inf}\big(\widehat{\nu}_{a,n}, \mu^\star - \delta\big) \leqslant \frac{1}{n} \ln\left(\frac{T}{Kn} \right) \right] \\ &\leqslant \sum_{n=1}^{f(T,K)} \mathbb{P}\left[\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn} \right) \right] = \mathbb{E}\left[\sum_{n=1}^{f(T,K)} \mathbb{1}_{\left\{ \sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn)) \right\}} \right] \\ &\leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{ \sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn)) \right\}} \right]. \end{aligned}$$

The last upper bound may seem crude but will be good enough for our purpose.

We may reinterpret

$$\mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right]$$

as the expected number of times a random walk with positive drift 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. In what follows, $\wedge$ denotes the minimum of two numbers. We define the first hitting time τ_a of the barrier, if it exists, as

$$\tau_a = \inf\left\{n \geqslant 1 :\ \sum_{i=1}^{n} Z_{a,i} > \ln\left(\frac{T}{Kn}\right)\right\} \wedge T\,.$$

The time τ_a is bounded by T and 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_a is met for some $1 \leqslant n \leqslant T$, i.e., whether or not the barrier is hit for $1 \leqslant n \leqslant T$, we get

$$S_4 \leqslant \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \leqslant \mathbb{E}[\tau_a] + \mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \tag{2.62}$$

where the sum from τ_a+1 to T is void thus null when $\tau_a = T$ (this is the case, in particular, when the barrier is hit for no $n \leqslant T$). We now state a lemma, in the spirit of Honda and Takemura [2015, Lemma 18], and will prove it later at the end of this section.

Lemma 2.8. *Let $(Z_i)_{i\geqslant 1}$ 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*

$$\tau \overset{def}{=} \inf\left\{n \geqslant 1 :\ \sum_{i=1}^{n} Z_i > \ln\left(\frac{T}{Kn}\right)\right\} \wedge T$$

and denote by W Lambert's function. Then, for all $T \geqslant K\mathrm{e}^{\alpha}$,

$$\mathbb{E}[\tau] \leqslant \frac{W(\alpha T/K) + \alpha + \ln 2}{\mathbb{E}[Z_1]}\,.$$

The random variables $Z_{a,i}$ have positive expectation $\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) > 0$ and are bounded by $\alpha = \ln\big(1/(1-\mu^\star)\big)$; indeed, since $X_{a,i} \geqslant 0$ and $\lambda_{a,\delta} \in [0,1]$, we have

$$\begin{aligned} Z_{a,i} = \ln\left(1 - \lambda_{a,\delta}\frac{X_{a,i} - (\mu^\star - \delta)}{1 - (\mu^\star - \delta)}\right) &\leqslant \ln\left(1 + \lambda_{a,\delta}\frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right) \\ &\leqslant \ln\left(1 + \frac{\mu^\star - \delta}{1 - (\mu^\star - \delta)}\right) = \ln\left(\frac{1}{1 - (\mu^\star - \delta)}\right) \\ &\leqslant \ln\left(\frac{1}{1-\mu^\star}\right) \overset{\text{def}}{=} \alpha\,. \end{aligned}$$

In addition, we imposed that $T > K/(1-\mu^\star) = K\mathrm{e}^{\alpha}$. Therefore, Lemma 2.8 applies and yields the bound

$$\begin{aligned} \mathbb{E}[\tau_a] &\leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \\ &\leqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - \delta/(1-\mu^\star)}\left(W\left(\frac{\ln\big(1/(1-\mu^\star)\big)}{K}\,T\right) + \ln\big(2/(1-\mu^\star)\big)\right) \end{aligned}$$

where the second inequality follows by the regularity inequality (2.19) on $\mathcal{K}_{\inf}$ (and the denominator therein is still positive thanks to our assumption on δ). All in all, we obtained the first part of the bound (2.60) and conclude the proof of the latter based on the decomposition (2.62) by showing that

$$\mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \leqslant \beta \stackrel{\text{def}}{=} 5 + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu_a,\mu^\star)^2/(8\gamma_\star)}}\,. \tag{2.63}$$

To that end, note that when $\tau_a < T$, we have by definition of τ_a,

$$\ln\left(\frac{T}{K\tau_a}\right) < \sum_{i=1}^{\tau_a} Z_{a,i}$$

The following implication thus holds for any $n \geqslant \tau_a$:

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \leqslant \ln\left(\frac{T}{K\tau_a}\right) \leqslant \sum_{i=1}^{\tau_a} Z_{a,i}\,. \tag{2.64}$$

Hence, in this case,

$$\sum_{i=1}^{n} Z_{a,i} \leqslant \ln\left(\frac{T}{Kn}\right) \qquad \text{implies} \qquad \sum_{i=\tau_a+1}^{n} Z_{a,i} < 0\,.$$

This, together with a breakdown according to the values of τ_a (note that the case $\tau_a = T$ does not contribute to the expectation) and the independence between $\{\tau_a = k\}$ and $Z_{a,k+1}, \dots, Z_{a,T}$, yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] = \mathbb{E}\left[\mathbb{1}_{\{\tau_a<T\}} \sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=1}^{n} Z_{a,i} \leqslant \ln(T/(Kn))\right\}}\right] \\
&\leqslant \mathbb{E}\left[\mathbb{1}_{\{\tau_a<T\}} \sum_{n=\tau_a+1}^{T} \mathbb{1}_{\left\{\sum_{i=\tau_a+1}^{n} Z_{a,i} < 0\right\}}\right] = \sum_{k=1}^{T-1} \mathbb{E}\left[\mathbb{1}_{\{\tau_a=k\}} \sum_{n=k+1}^{T} \mathbb{1}_{\left\{\sum_{i=k+1}^{n} Z_{a,i} < 0\right\}}\right] \\
&= \sum_{k=1}^{T-1} \sum_{n=k+1}^{T} \mathbb{P}[\tau_a = k]\ \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} < 0\right] \\
&= \sum_{k=1}^{T-1} \mathbb{P}[\tau_a = k] \left(\underbrace{\sum_{n=k+1}^{T} \mathbb{P}\left[\sum_{i=k+1}^{n} Z_{a,i} < 0\right]}_{\text{we show below } \leqslant \beta, \text{ see (2.67)}}\right) \leqslant \beta
\end{aligned} \tag{2.65}$$

where β was defined in (2.63).

Indeed, we resort to Remark 5 of Section 2.B.4, for the $n-k$ variables $Z_{a,k+1}, \dots, Z_{a,n}$ and $x = 0$; we legitimately do so as $\mu^\star - \delta > \mu_a$ by the imposed condition $\delta < \Delta_a/2$. Thus, denoting

$$\gamma_{\star,\delta} = \frac{1}{\sqrt{1-(\mu^\star-\delta)}}\left(16\mathrm{e}^{-2} + \ln^2\left(\frac{1}{1-(\mu^\star-\delta)}\right)\right) \leqslant \gamma_\star$$

we have

$$\begin{aligned}\mathbb{P}\bigg[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\bigg] &\leqslant \max\bigg\{\mathrm{e}^{-(n-k)/4},\ \exp\bigg(-\frac{n-k}{2\gamma_{\star,\delta}}\Big(\mathcal{K}_{\inf}(\nu_a,\mu^\star-\delta)\Big)^2\bigg)\bigg\}\\ &\leqslant \mathrm{e}^{-(n-k)/4} + \exp\bigg(-\frac{n-k}{2\gamma_\star}\Big(\mathcal{K}_{\inf}(\nu_a,\mu^\star-\delta)\Big)^2\bigg)\\ &\leqslant \mathrm{e}^{-(n-k)/4} + \mathrm{e}^{-(n-k)\mathcal{K}_{\inf}(\nu,\mu^\star)^2/(8\gamma_\star)}\end{aligned}$$

where the third inequality follows from (2.19) and from the imposed condition $\delta \leqslant (1-\mu^\star)\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)/2$:

$$\mathcal{K}_{\inf}(\nu_a,\mu^\star-\delta) \geqslant \mathcal{K}_{\inf}(\nu_a,\mu^\star) - \frac{\delta}{1-\mu^\star} \geqslant \frac{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}{2}\,. \tag{2.66}$$

We finally get, after summation over $n = k+1,\dots,T$,

$$\sum_{n=k+1}^{T}\mathbb{P}\bigg[\sum_{i=k+1}^{n} Z_{a,i} \leqslant 0\bigg] \leqslant \underbrace{\frac{1}{1-\mathrm{e}^{-1/4}}}_{\leqslant 5} + \frac{1}{1-\mathrm{e}^{-\mathcal{K}_{\inf}(\nu_a,\mu^\star)^2/(8\gamma_\star)}}\,, \tag{2.67}$$

which is the inequality claimed in (2.65).

It only remains to prove Lemma 2.8.

Proof of Lemma 2.8. This lemma 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 complete proof, with no significant additional merit compared to the original 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 exactly 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 $\ln\big(T/(Kn)\big)$ 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, that

$$\sum_{i=1}^{\tau-1} Z_i \leqslant \ln\bigg(\frac{T}{K(\tau-1)}\bigg)$$

and thus

$$\sum_{i=1}^{\tau-1} Z_i + Z_\tau \leqslant \ln\bigg(\frac{T}{K(\tau-1)}\bigg) + \alpha = \ln\bigg(\frac{T}{K\tau}\bigg) + \ln\bigg(\frac{\tau}{\tau-1}\bigg) + \alpha \leqslant \ln\bigg(\frac{T}{K\tau}\bigg) + \ln 2 + \alpha\,.$$

In addition, when $\tau < T/K$, and again by definition of τ,

$$\ln\left(\frac{T}{K\tau}\right) < \sum_{i=1}^{\tau} Z_i \leqslant \tau\alpha \qquad \text{thus} \qquad 0 < \frac{T}{K\tau}\ln\left(\frac{T}{K\tau}\right) \leqslant \frac{T\alpha}{K}\,.$$

Applying the increasing function W to both sides of the latter inequality, we get, when $\tau < T/K$,

$$\ln\left(\frac{T}{K\tau}\right) \leqslant W\left(\frac{T\alpha}{K}\right).$$

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

$$\mathbb{E}[\tau]\,\mathbb{E}[Z_1] = \mathbb{E}\left[\sum_{i=1}^{\tau} Z_i\right] \leqslant W\left(\frac{T\alpha}{K}\right) + \ln 2 + \alpha$$

which concludes the proof. □
□

2.D. Proof of the variational formula (Lemma 2.3)

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 that does not rely on discrete approximations of general distributions.

But before we do so, we explain why it is natural to expect to rewrite $\mathcal{K}_{\inf}$, which is an infimum, as a maximum. Indeed, given that Kullback-Leibler divergences are given by a supremum, $\mathcal{K}_{\inf}$ 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

$$\mathrm{KL}(\nu,\nu') = \sup\Big\{\mathbb{E}_\nu[Y] - \ln \mathbb{E}_{\nu'}\big[\mathrm{e}^Y\big] :\ Y \text{ s.t. } \mathbb{E}_{\nu'}[\mathrm{e}^Y] < +\infty\Big\} \tag{2.68}$$

where (only in the next few lines) we index the expectation with respect to the assumed distribution of the random variable Y. In particular, denoting by X the identity over $[0,1]$ and considering, for $\lambda \in [0,1]$, the variables bounded from above

$$Y_\lambda = \ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right) \leqslant \ln\left(1 + \frac{\lambda\mu}{1-\mu}\right)$$

we have, for any probability measure ν' such that $\mathrm{E}(\nu') > \mu$:

$$\ln \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big] = \ln\left(\mathbb{E}_{\nu'}\left[1 - \lambda\frac{X-\mu}{1-\mu}\right]\right) = \ln\left(1 - \lambda\frac{\mathrm{E}(\nu')-\mu}{1-\mu}\right) \leqslant 0\,.$$

Hence, for these distributions ν',

$$\mathrm{KL}(\nu,\nu') \geqslant \sup_{\lambda\in[0,1]}\Big\{\mathbb{E}_\nu[Y_\lambda] - \ln \mathbb{E}_{\nu'}\big[\mathrm{e}^{Y_\lambda}\big]\Big\} \geqslant \sup_{\lambda\in[0,1]} \mathbb{E}_\nu\left[\ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right]$$

and by taking the infimum over all distributions ν' with $\mathrm{E}(\nu') > \mu$:

$$\mathcal{K}_{\inf}(\nu,\mu) \geqslant \sup_{\lambda\in[0,1]} \mathbb{E}_\nu\left[\ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right]. \tag{2.69}$$

Outline. We now only need to prove the converse inequality to get the rewriting (2.50) of Lemma 2.3, which we will do in Section 2.D.2. Before that, in Section 2.D.1, we prove the second statement of Lemma 2.3 together with several useful facts for the proof provided in Section 2.D.2, including the fact that the supremum in the right-hand side of (2.69) is achieved. We conclude in Section 2.D.3 with an alternative (sketch of) proof of the inequality (2.69), not relying on the variational formula (2.68) for the Kullback-Leibler divergences.

2.D.1. A function study

Let X denote a random variable with distribution $\nu \in \mathcal{P}[0,1]$. We recall that $\mu \in (0,1)$. The following function is well defined:

$$H : \lambda \in [0,1] \longmapsto \mathbb{E}\left[\ln\left(1 - \lambda\frac{X-\mu}{1-\mu}\right)\right] \in \mathbb{R}\cup\{-\infty\}\,.$$

Indeed, since $X \in [0,1]$, the random variable $\ln\big(1-\lambda(X-\mu)/(1-\mu)\big)$ is bounded from above by $\ln\big(1+\lambda\mu/(1-\mu)\big)$. Hence, H is well defined. For $\lambda \in [0,1)$, the considered random variable is bounded from below by $\ln(1-\lambda)$, hence H takes finite values. For $\lambda = 1$, we possibly have that $H(1)$ equals $-\infty$ (this is the case in particular when $\nu\{1\} > 0$).

We begin by a study of the function H.

Lemma 2.9. *The function H is continuous and strictly concave on $[0,1]$, differentiable at least on $[0,1)$, and its derivative $H'(1)$ can be defined at 1, with $H'(1) \in \mathbb{R} \cup \{-\infty\}$. We have the closed-form expression: for all $\lambda \in [0,1]$,*

$$H'(\lambda) = -\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right] = \frac{1}{\lambda}\left(1-\mathbb{E}\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right). \tag{2.70}$$

It reaches a unique maximum over $[0,1]$, denoted by $\lambda^\star$,

$$\arg\max_{0\leqslant\lambda\leqslant 1} H(\lambda) = \{\lambda^\star\}$$

at which $H'(\lambda^\star) = 0$ if $\lambda^\star \in [0,1)$ and $H'(\lambda^\star) \geqslant 0$ if $\lambda^\star = 1$.

Moreover, under the additional condition $\mathrm{E}(\nu) < \mu$,

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = 1 \quad \text{if } \lambda^\star \in [0,1) \qquad \text{and} \qquad \mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = \mathbb{E}\left[\frac{1-\mu}{1-X}\right] \leqslant 1 \quad \text{if } \lambda^\star = 1$$

where we have in particular $\nu\{1\} = 0$ in the latter case $\lambda^\star = 1$.

Note that $\mathcal{K}_{\inf}(\nu,\mu) = 0$ when $\mu \leqslant \mathrm{E}(\nu)$. In this case, necessarily $\lambda^\star = 0$ (there is a unique maximum) and we still have

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star\frac{X-\mu}{1-\mu}}\right] = 1\,.$$

This concludes the proof of the statement (2.51) of Lemma 2.3.

Proof. For the continuity of H, we note that the discussion before the statement of the lemma entails that the random variables $\ln\big(1-\lambda(X-\mu)/(1-\mu)\big)$ are uniformly bounded on ranges of the form $[0,\lambda_0]$ for $\lambda_0 < 1$. By a standard continuity theorem under the integral sign, this proves that H is continuous on $[0,1)$. For the continuity at 1, we separate the $H(\lambda)$ and $H(1)$ into two pieces, for which monotone convergences take place:

$$\lim_{\lambda\to 1}\mathbb{E}\left[\ln\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X\in[0,\mu]\}}\right] = \mathbb{E}\left[\ln\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X\in[0,\mu]\}}\right]$$
$$\lim_{\lambda\to 1}\mathbb{E}\left[\ln\left(1-\lambda\frac{X-\mu}{1-\mu}\right)\mathbb{1}_{\{X\in(\mu,1]\}}\right] = \mathbb{E}\left[\ln\left(\frac{1-X}{1-\mu}\right)\mathbb{1}_{\{X\in(\mu,1]\}}\right]$$

where the first expectation is finite (but the second may equal $-\infty$).

The strict concavity of H on $[0,1]$ follows from the one of ln on $(0,1]$ and from the continuity of H on $[0,1]$.

For $\lambda \in [0,1)$, we get, by legitimately differentiating under the expectation,

$$H'(\lambda) = -\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu}\right)\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right] = \frac{1}{\lambda}\left(1-\mathbb{E}\left[\frac{1}{1-\lambda\frac{X-\mu}{1-\mu}}\right]\right).$$

Indeed as long as $\lambda < 1$, the random variables in the expectations are uniformly bounded on ranges of the form $[0, \lambda_0]$ for $\lambda_0 < 1$, so that we may invoke a standard differentiation theorem under the integral sign. A similar argument of double monotone convergences as above shows that $H'(\lambda)$ has a limit value as $\lambda \to 1$, with

$$\lim_{\lambda \to 1} H'(\lambda) = -\mathbb{E}\left[\frac{X-\mu}{1-X}\right].$$

By a standard limit theorem on derivatives, when the above value is finite, H is differentiable at 1 and $H'(1)$ equals the limit above; otherwise, H is not differentiable at 1 but we still denote $H'(1) = -\infty$.

Since H is strictly concave on $[0,1]$ and continuous, it reaches its maximum exactly once on $[0,1]$. Now, under the condition $\mu < \mathrm{E}(\nu) < 1$, we have

$$H'(0) = -\frac{\mathrm{E}(\nu) - \mu}{1-\mu} > 0\,.$$

As H is concave, H' is decreasing: either $H'(1) \geqslant 0$ and H reaches its maximum at $\lambda^\star = 1$, or $H'(1) < 0$ and H reaches its maximum on the open interval $(0,1)$. It may be proved (by a standard continuity theorem under the integral sign) that H' is continuous on $[0,1)$, that is, that H is continuously differentiable on $[0,1)$. In the case $H'(1) < 0$, the derivative at the maximum therefore satisfies $H'(\lambda^\star) = 0$. Substituting the expression for $H'(\lambda^\star)$ concludes the proof. □

2.D.2. Proof of $\leqslant$ in the equality (2.50)

We keep the notation introduced in the previous section. To prove this inequality, by the rewriting of $\mathcal{K}_{\inf}(\nu,\mu)$ stated in Corollary 2.2, it is enough to show that there exists a probability measure ν' on $[0,1]$ such that $\mathrm{E}(\nu') \geqslant \mu$ and $\nu \ll \nu'$ and

$$\mathrm{KL}(\nu,\nu') \leqslant \mathbb{E}\left[\ln\left(1 - \lambda^\star \frac{X-\mu}{1-\mu}\,.\right)\right] \tag{2.71}$$

Given the definition of the KL divergence, it suffices to find a probability measure ν' on $[0,1]$ such that $\mathrm{E}(\nu') \geqslant \mu$ and $\nu \ll \nu'$ and

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = 1 - \lambda^\star \frac{x-\mu}{1-\mu} \qquad \nu\text{–a.s.}\,. \tag{2.72}$$

It can be shown (proof omitted as this statement is only given to explain the intuition behind the proof) that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'} > 0 \quad \nu\text{–a.s.} \qquad \text{with} \qquad \frac{\mathrm{d}\nu'_{\mathrm{ac}}}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)^{-1} \quad \nu\text{–a.s.} \tag{2.73}$$

where ν'_{ac} denotes the absolute part of ν' with respect to ν. This is why we introduce the measure ν' on $[0,1]$ defined by

$$\mathrm{d}\nu'(x) = \underbrace{\frac{1}{1 - \lambda^\star \frac{x-\mu}{1-\mu}}}_{\geqslant 0}\,\mathrm{d}\nu(x) + \left(1 - \mathbb{E}\left[\frac{1}{1 - \lambda^\star \frac{X-\mu}{1-\mu}}\right]\right)\mathrm{d}\delta_1(x)$$

where δ_1 denotes the Dirac point-mass distribution at 1 and where X denotes a random variable with distribution ν. The measure ν' is a probability measure as by Lemma 2.9,

$$\mathbb{E}\left[\frac{1}{1 - \lambda^\star \frac{X-\mu}{1-\mu}}\right] \leqslant 1\,.$$

Now, we show first that $\nu \ll \nu'$ with the density (2.72). We do so by distinguishing two cases. If $\lambda^\star \in [0,1)$, then by the last statement of Lemma 2.9, the probability measure ν' is actually defined by

$$\mathrm{d}\nu'(x) = \underbrace{\frac{1}{1-\lambda^\star \frac{x-\mu}{1-\mu}}}_{>0} \, \mathrm{d}\nu(x)$$

and the strict positivity underlined in the equality above ensures the desired result by a standard theorem on Radon-Nikodym derivatives. In that case, ν and ν' are actually equivalent measures: $\nu \ll \nu'$ and $\nu' \ll \nu$. If $\lambda^\star = 1$, then again by Lemma 2.9, we know that ν does not put any probability mass at 1. The strict positivity of $f(x) = 1-(x-\mu)/(1-\mu)$ on $[0,1)$ and the fact that $\nu\{1\} = 0$ ensure the first equality below: for all Borel sets A of $[0,1]$,

$$\nu(A) = \int \mathbb{1}_A \, f \frac{1}{f} \mathrm{d}\nu = \int \mathbb{1}_A \, f \left(\frac{1}{f} \mathrm{d}\nu + r \mathrm{d}\delta_1 \right) = \int \mathbb{1}_A \, f \mathrm{d}\nu'$$

while the second equality follows from $f(1) = 0$ and the third equality is by definition of ν'. Put differently, $\nu \ll \nu'$ with the density f claimed in (2.72). In that case, $\nu \ll \nu'$ but ν' is not necessarily absolutely continuous with respect to ν.

We conclude this proof by showing that $\mathrm{E}(\nu') \geqslant \mu$. We recall that Lemma 2.9 indicates that

$$\mathbb{E}\left[\left(\frac{X-\mu}{1-\mu} \right) \frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] = -H'(\lambda^\star)$$

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] = 1 - \lambda^\star \, H'(\lambda^\star)$$

where X denotes a random variable with distribution ν and where both expectations are well defined (possibly with values $+\infty$ when $\lambda^\star = 1$). Therefore,

$$\begin{aligned}
\mathrm{E}(\nu') &= \overbrace{\mathbb{E}\left[\frac{X}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right]}^{\text{“}\nu \text{ part of } \nu'\text{”}} + \overbrace{\left(1 - \mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] \right)}^{\text{“}\delta_1 \text{ part of } \nu'\text{”}} \\
&= (1-\mu) \, \mathbb{E}\left[\left(\frac{X-\mu}{1-\mu} \right) \frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] + \mu \, \mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] + \left(1 - \mathbb{E}\left[\frac{1}{1-\lambda^\star \frac{X-\mu}{1-\mu}} \right] \right) \\
&= -(1-\mu) \, H'(\lambda^\star) + \mu\big(1 - \lambda^\star \, H'(\lambda^\star)\big) + \lambda^\star \, H'(\lambda^\star) = \mu - \big((1-\mu)\,(1-\lambda^\star)\, H'(\lambda^\star)\big)
\end{aligned}$$

where the first equality is justified in the case $\lambda^\star = 1$ by the same arguments of monotone convergence as in the proof of Lemma 2.9. All in all, we have $\mathrm{E}(\nu') \geqslant \mu$ as desired if and only if $(1-\lambda^\star)\, H'(\lambda^\star) \leqslant 0$. This is the case as we actually have $(1-\lambda^\star)\, H'(\lambda^\star) = 0$ in all cases, i.e., whether $\lambda^\star = 1$ or $\lambda^\star \in [0,1)$.

2.D.3. Alternative proof of $\geqslant$ in the equality (2.50)

We use the notation of Sections 2.D.1 and 2.D.2 and prove the desired inequality (2.69), that is, the $\geqslant$ part of the equality (2.50), without resorting to the variational formula (2.68) for the Kullback-Leibler divergences. Actually, we only provide a sketch of proof and omit proofs of some facts about Radon-Nikodym derivatives.

Let $\nu'' \in \mathcal{P}[0,1]$ be such that $\mathrm{E}(\nu'') > \mu$ and $\nu \ll \nu''$; with no loss of generality, we assume that $\mathrm{KL}(\nu, \nu'') < +\infty$. By definition of ν', the divergence $\mathrm{KL}(\nu, \nu')$ equals the maximum of the

continuous function H over $[0,1]$ and therefore also satisfies $\mathrm{KL}(\nu,\nu') < +\infty$. We denote by $\mathbb{L}_1(\nu)$ the set of ν–integrable random variables. That these divergences are finite means that

$$\left|\ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right| \in \mathbb{L}_1(\nu) \qquad \text{and} \qquad \left|\ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu''}\right| \in \mathbb{L}_1(\nu)\,.$$

Hence,

$$\mathrm{KL}(\nu,\nu'') - \mathrm{KL}(\nu,\nu') = -\int \left(\ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu'} - \ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu''}\right)\mathrm{d}\nu\,.$$

Now, by (2.72),

$$\ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = \ln\left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right) \qquad \nu\text{–a.s.}$$

and by (2.73),

$$-\ln \frac{\mathrm{d}\nu}{\mathrm{d}\nu''} = \ln \frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x) \qquad \nu\text{–a.s.}$$

so that

$$\begin{aligned}
\mathrm{KL}(\nu,\nu'') - \mathrm{KL}(\nu,\nu') &= -\int \ln\left(\left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right)\frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x)\right)\mathrm{d}\nu(x) \\
&\geqslant -\ln\left(\int \underbrace{\left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right)}_{\geqslant 0} \underbrace{\frac{\mathrm{d}\nu''_{\mathrm{ac}}}{\mathrm{d}\nu}(x)\,\mathrm{d}\nu(x)}_{\mathrm{d}\nu''_{\mathrm{ac}}(x)}\right) \\
&\geqslant -\ln\left(\underbrace{\int \left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right)\mathrm{d}\nu''(x)}_{\leqslant 1 \text{ as } \mathrm{E}(\nu'')>\mu}\right) \geqslant 0
\end{aligned}$$

where Jensen's inequality provided the first inequality, while the second one followed by increasing the integral in the logarithm. Taking the infimum over distributions $\nu'' \in \mathcal{P}[0,1]$ with $\mathrm{E}(\nu'') > \mu$ and $\nu \ll \nu''$ and $\mathrm{KL}(\nu,\nu'') < +\infty$, we proved

$$\mathcal{K}_{\inf}(\nu,\mu) - \mathrm{KL}(\nu,\nu') \geqslant 0$$

which was the desired result.

Chapter 3.

Polynomial cost of adaptation for $\mathcal{X}$-armed bandits

Abstract

In the context of stochastic continuum-armed bandits, we present an algorithm that adapts to the unknown smoothness of the objective function. We exhibit and compute a *polynomial cost of adaptation* to the Hölder regularity for regret minimization. To do this, we first reconsider the recent lower bound of Locatelli and Carpentier [2018], and define and characterize admissible rate functions. Our new algorithm matches any of these minimal rate functions. We provide a finite-time analysis and a thorough discussion about asymptotic optimality.

This work led to the publication Hadiji [2019] at the conference Neural Information Processing Systems (Neurips 2019).

Contents

3.1. Introduction

Multi-armed bandits are a well-known sequential learning problem. When the number of available decisions is large, some assumptions on the environment have to be made. In a vast line of work (see the literature discussion in Section 3.1.1), these assumptions show up as nonparametric regularity conditions on the mean-payoff function. If this function is Hölder continuous with constant L and exponent α, and if the values of L and α are given to the player, then natural strategies can ensure that the regret is upper bounded by

$$L^{1/(2\alpha+1)}T^{(\alpha+1)/(2\alpha+1)}\,. \tag{3.1}$$

Of course, assuming that the player knows α and L is often not realistic. Thus the need for *adaptive* methods, that are agnostic with respect to the true regularity of the mean-payoff function. Unfortunately, Locatelli and Carpentier [2018] recently showed that full adaptation is impossible, and that no algorithm can enjoy the same minimax guarantees as when the regularity is given to the player. We persevere and address the question:

What can the player achieve when the true regularity is completely unknown?

A polynomial cost of adaptation In statistics, minimax adaptation for nonparametric function estimation is a deep and active research domain. In many contexts, sharp adaptation is possible; often, an additional logarithmic factor in the error has to be paid when the regularity is unknown: this is known as the *cost of adaptation*. See e.g., Lepskii [1991], Birgé and Massart [1995], Massart [2007] for adaptive methods, and Cai [2012] for a detailed survey of the topic. Under some more exotic assumptions —see e.g., Example 3 of Cai and Low [2005] — adapting is significantly harder: there may be a *polynomial cost of adaptation*.

In this chapter, we show that in the sequential setting of multi-armed bandits, the necessary exploration forces a similar phenomenon, and we exhibit this polynomial cost of adaptation. To do so, we revisit the lower bounds of Locatelli and Carpentier [2018], and design a new algorithm that matches these lower bounds.

As a representative example of our results, our algorithm can achieve, without the knowledge of α and L, an unimprovable (up to logarithmic factors) regret bound of order

$$L^{1/(1+\alpha)}T^{(\alpha+2)/(2\alpha+2)}\,. \tag{3.2}$$

3.1.1. Related work

Continuum-armed bandits Continuum-armed bandits, with nonparametric regularity assumptions, were introduced by Agrawal [1995a]. Kleinberg [2004] established the minimax rates in the Hölder setting and introduced the CAB1 algorithm. Auer et al. [2007] studied the problem with additional regularity assumptions under which the minimax rates are improved. Via different roads, Bubeck et al. [2011b] and Kleinberg et al. [2019] explored further generalizations of these types of regularity, namely the zooming dimension and the near-optimality dimension. Bull [2015] exhibited an algorithm that essentially adapts to some cases when the near-optimality dimension is zero.

In all these articles, the mean-payoff function needs to satisfy simultaneously two sets of regularity conditions. The first type is a usual Hölder condition, which ensures that the function does not vary too much around (one of) its maxima. The second type is a "margin condition" that lower bounds the number of very suboptimal arms; in the literature these are defined in many technically different ways. Adapting to the margin conditions is often possible when the Hölder regularity is known. However, all these algorithms require some prior knowledge about the Hölder regularity.

In this chapter, we focus on the problem of adapting to Hölder regularity. Accordingly, we call *adaptive* the algorithms that assume no knowledge of the Hölder exponent nor of the Lipschitz constant.

Adaptation for cumulative regret Bubeck et al. [2011c] introduced the problem of adaptation, and adapted to the Lipschitz constant under extra requirements. An important step was made in Locatelli and Carpentier [2018], where it is shown that adaptation at the classical minimax rates is impossible. In the same article, the authors exhibited some conditions under which full adaptation is achievable, e.g., with knowledge of the value of the maximum, or when the near-optimality dimension is zero.

Other settings For simple regret, the objections against adaptation do not hold, as the objective does not penalize exploration. Adaptation up to polylog factors is done with various (meta-)algorithms. Locatelli and Carpentier [2018] sketch out an aggregation approach inspired by Lepski's method, while Valko et al. [2013], Grill et al. [2015], Shang et al. [2019] describe cross-validation methods thanks to which they adapt to the near-optimality dimension with unknown smoothness. As it turns out, this last approach yields clean results with our smoothness assumptions; we write the details in Appendix 3.C.

There is also an important segment of the literature on nonparametric estimation that is devoted to the estimation of the maximum of a smooth function. For this problem, the optimal asymptotic rates of estimation have been derived to remarkable preicison Lepskii [1994].

Recently, Krishnamurthy et al. [2019] studied continuum-armed contextual bandits and used a sophisticated aggregation scheme to derive an algorithm that adapts to the Lipschitz constant when $L \geqslant 1$.

3.1.2. Contributions and outline

In this chapter, we fully compute the cost of adaptation for bandits with Hölder regularity. In Section 3.2 we discuss the adaptive (and nonadaptive) lower bounds. We take an asymptotic stance in order to precisely define the objective of adaptation. Doing so, we uncover a family of noncomparable lower bounds for adaptive algorithms (Theorem 3.1), and define the corresponding notion of optimality: admissibility.

Section 3.3 contains our main contribution: an admissible adaptive algorithm. We first recall the CAB1 algorithm, which is nonadaptive minimax, and use it as a building block for our new algorithm (Subsection 3.3.1). This algorithm works in a regime-based fashion. Between successive regimes of doubling lengths, we reset the algorithm and use a new discretization with fewer arms. In order to carry information between the different stages, we use CAB1 in a clever way: besides partitioning the arm space, we add summaries of previous regimes by allowing the algorithm to play according to the empirical distributions of past plays. This is formally described in Subsection 3.3.2.

A salient difference with all previous approaches is that we zoom out by using fewer and fewer arms. To our knowledge, this is unique, as all other algorithms for bandits zoom in in a way that crucially depends on the regularity parameters. Another important feature of our analysis is that we adapt both to the Hölder exponent α and to the Lipschitz constant L. On a technical level, this is thanks to the fact that we do not explicitly choose a grid of regularity parameters, which means that we implicitly handle all values (L, α) simultaneously.

We first give a regret bound in the known horizon case (Subsection 3.3.2), then we provide an anytime version and we show that they match the lower bounds of adaptation (Subsection 3.3.4). Finally Section 3.4 provides the proof of our main regret bound.

3.2. Setup, preliminary discussion

3.2.1. Notation and known results

Let us reintroduce briefly the standard bandit terminology. We consider the arm space $\mathcal{X} = [0,1]$. The environment sets a reward function $f : \mathcal{X} \to [0,1]$. At each time step t, the player chooses an arm $X_t \in \mathcal{X}$, and the environment then displays a reward Y_t such that $\mathbb{E}[Y_t \mid X_t] = f(X_t)$, independently from the past. We assume that the variables $Y_t - f(X_t)$ are $(1/4)$-subgaussian conditionnally on X_t; this is satisfied if the payoffs are bounded in $[0,1]$ by Hoeffding's lemma.

The objective of the player is to find a strategy that minimizes her *expected cumulative (pseudo-)regret*. If $M(f)$ denotes the maximum value of f, the regret at time T is defined as

$$\overline{R}_T = TM(f) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = TM(f) - \mathbb{E}\left[\sum_{t=1}^{T} f(X_t)\right] . \tag{3.3}$$

In this chapter, we assume that the function f satisfies a Hölder assumption around one of its maxima:

Definition 3.1. *For $\alpha > 0$ and $L > 0$, we denote by $\mathcal{H}(L,\alpha)$ the set of functions that satisfy*

$$\exists\, x^\star \in [0,1] \;\; s.t. \;\; f(x^\star) = M(f) \;\; and \;\; \forall\, x \in [0,1] \quad |f(x^\star) - f(x)| \leqslant L\,|x^\star - x|^\alpha \; . \tag{3.4}$$

Note that this assumption is a lot weaker than the standard Hölder assumption, and that functions satisfying this condition could be vastly irregular.

We are interested in minimax rates of regret when the mean-payoff function f belongs to these Hölder-type classes, i.e., the quantity $\inf_{\text{algorithms}} \sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T$.

MOSS Throughout this chapter, we exploit discretization arguments and use a minimax optimal algorithm for finite-armed bandits: MOSS, from Audibert and Bubeck [2009]. When run for T rounds on a K-armed bandit problem with $(1/4)$-subgaussian noise, and when $T \geqslant K$, its regret is upper-bounded by $18\sqrt{KT}$ (the improved constant is from Garivier et al. [2018]).

Non-adaptive minimax rates When the regularity is given to the player, for any α, L and T:

$$0.001\, L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)} \leqslant \inf_{\text{algorithms}} \sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 28\, L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)} . \tag{3.5}$$

This is well-known since Kleinberg [2004]. For completeness, we recall how to derive the upper bound in Section 3.3.1, and the lower bound in Section 3.2.2.

3.2.2. Lower bounds: adaptation *at usual rates* is not possible

Locatelli and Carpentier [2018] prove a version of the following theorem; see our reshuffled and slightly improved proof in Appendix 3.D.

Theorem (Variation on Th.3 from Locatelli and Carpentier [2018]). *Let $B > 0$ be a positive number. Let $\alpha, \gamma > 0$ and $L, \ell > 0$ be regularity parameters that satify $\alpha \leqslant \gamma$ and $L \geqslant \ell$.*

Assume moreover that $2^{-3}\, 12^\alpha B^{-1} \leqslant L \leqslant \ell^{1+\alpha}\, T^{\alpha/2}\, 2^{(1+\alpha)(8-2\gamma)}$. If an algorithm is such that $\sup_{f\in\mathcal{H}(\ell,\gamma)} \overline{R}_T \leqslant B$, then the regret of this algorithm is lower bounded on $\mathcal{H}(L,\alpha)$:

$$\sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T \geqslant 2^{-10}\, T L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)} . \tag{3.6}$$

Remark (Bibliographical note). *Locatelli and Carpentier [2018] consider a more general setting where additional margin conditions are exploited. In our setting, we slightly improve their result by dealing with the dependence on the Lipschitz constant, and by removing a requirement on B.*

In a different context, Krishnamurthy et al. [2019] show a variation of this bound where the Lipschitz constant is considered, but only in the case where $\alpha = \gamma = 1$, for $\ell = 1$ and $L \geqslant 1$.

As explained in Locatelli and Carpentier [2018] this forbids adaptation at the usual minimax rates over two regularity classes; we recall how in the paragraph that follows Theorem 3.1. However this is not the end of the story, as one naturally wonders what is the best the player can do.

To further investigate this question, we discuss it asymptotically by considering the rates at which the minimax regret goes to infinity, therefore focusing on the dependence on T. Our main results are completely nonasymptotic, yet we feel the asymptotic analysis of optimality is clearer.

Definition 3.2. *Let $\theta : [0,1] \to [0,1]$ denote a nonincreasing function. We say an algorithm* achieves adaptive rates θ *if*

$$\forall\, \varepsilon > 0\,,\ \forall\, \alpha,\, L\, > 0\,, \quad \limsup_{T\to\infty} \frac{\sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T}{T^{\theta(\alpha)+\varepsilon}} < +\infty\,.$$

We include the ε in the definition in order to neglect the potential logarithmic factors.

As rate functions are not always comparable for pointwise order, the good notion of optimality is the standard statistical notion of *admissibility* (akin to "Pareto optimality" for game-theorists).

Definition 3.3. *A rate function is said to be* admissible *if it is achieved by some algorithm, and if no other algorithm achieves stricly smaller rates for pointwise order. An algorithm is admissible if it achieves an admissible rate function.*

We recall that a function θ' is stricly smaller than θ for pointwise order if $\theta'(\alpha) \leqslant \theta(\alpha)$ for all α and $\theta'(\alpha_0) < \theta(\alpha_0)$ for at least one value of α_0.

It turns out we can fully characterize the admissible rate functions by inspecting the lower bounds (3.6).

Theorem 3.1. *The admissible rate functions are exactly the family*

$$\theta_m : \alpha \mapsto \max\left(m\,,\ 1 - m\frac{\alpha}{\alpha+1}\right)\,, \quad m \in [1/2, 1]\,. \tag{3.7}$$

This theorem contains two assertions. The lower bound side states that no smaller rate function may be achieved by any algorithm. This side is derived from an asymptotic rewording of lower bound (3.6), see Proposition 3.1 stated below. The proof is done through a careful inspection of the functional inequation defining the lower bound. The second statement is that the θ_m's are indeed achieved by an algorithm, which is the subject of Section 3.3.2.

Proof. First of all, by Corollary 3.2, the appropriately tuned MeDZO may achieve all the $\theta'_m s$. Thus we are left to prove the lower bound side, i.e., that all the admissible rate functions belong to the family θ_m.

The best way to see this is to first notice that for θ nonincreasing and positive, the inequation in Proposition 3.1 is equivalent to

$$\forall\, \alpha > 0\,, \quad \theta(\alpha) \geqslant 1 - \theta(\infty)\frac{\alpha}{\alpha+1}\,. \tag{3.8}$$

Notice that taking $\gamma = +\infty$ is always valid in what follows, as θ is assumed to be nonincreasing and lower bounded by $1/2$. Now if θ satisfies (3.9), then it satisfies (3.8) by taking $\gamma = +\infty$. For the converse, consider $\alpha \leqslant \gamma$, then $\theta(\gamma) \geqslant \theta(\infty)$, thus $1 - \theta(\infty)\alpha/(\alpha+1) \geqslant 1 - \theta(\gamma)\alpha/(\alpha+1)$.

Now consider an admissible θ. Since θ is achieved by some algorithm, by Proposition 3.1 and the remark above, it satisfies Eq. (3.8). As θ is nonincreasing, and by Eq. (3.8), we have $\theta(\alpha) \geqslant \theta(\infty)$ and $\theta(\alpha) \geqslant 1 - \theta(\infty)\alpha/(\alpha+1)$. In other words, $\theta \geqslant \theta_{m_\theta}$, where $m_\theta = \theta(\infty) \in [1/2, 1]$. By the admissibility of θ, this implies that $\theta = \theta_{m_\theta}$. □

Figure 3.1 illustrates how these admissible rates compare to each other, and to the usual minimax rates.

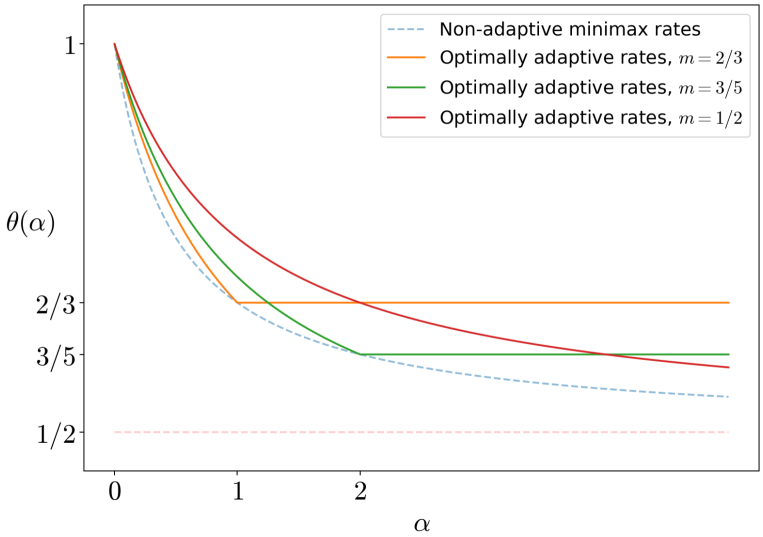


Figure 3.1.: The lower bounds on adaptive rates: plots of the admissible rate functions $\alpha \mapsto \theta_m(\alpha)$. If an algorithm has regret of order $\mathcal{O}\big(T^{\theta(\alpha)}\big)$, then θ is everywhere above one of these curves.

In particular, we see that reaching the nonadaptive minimax rates for multiple values of α is impossible. Moreover, at $m = (\gamma+1)/(2\gamma+1)$, we have $\theta_m(\gamma) = (\gamma+1)/(2\gamma+1)$, which is the usual minimax rate (3.1) when γ is known. This yields an alternative parameterization of the family θ_m: one may choose to parameterize the functions either by their value at infinity $m \in [1/2, 1]$, or by the only point $\gamma \in [0, +\infty]$ at which they coincide with the usual minimax rates function (3.1).

Proposition 3.1. *Assume an algorithm achieves adaptive rates $\theta : [0, +\infty) \to [0, 1]$. Then θ satisfies the functional inequation*

$$\forall\, \gamma > 0\,, \quad \forall\, \alpha \leqslant \gamma\,, \quad \theta(\alpha) \geqslant 1 - \theta(\gamma)\frac{\alpha}{\alpha+1}\,. \tag{3.9}$$

Proof. Choose α, γ such that $\alpha \leqslant \gamma$, and $\varepsilon > 0$. Set $L > 0$. There exist constants c_1 and c_2 (depending on L, α, γ and ε) such that for T large enough,

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{R}_T \leqslant c_1 T^{\theta(\alpha)+\varepsilon} \quad \text{and} \quad \sup_{f \in \mathcal{H}(L,\gamma)} \overline{R}_T \leqslant c_2 T^{\theta(\gamma)+\varepsilon}\,.$$

Moreover, for T large enough, the assumptions for lower bound (3.6) hold. Hence applying the lower bound with $B = c_2 T^{\theta(\gamma)+\varepsilon}$, for some constant c_3:

$$c_1 T^{\theta(\alpha)+\varepsilon} \geqslant 0.0001\, T\big(c_2 T^{\theta(\gamma)+\varepsilon}\big)^{-\alpha/(\alpha+1)} \geqslant c_3\, T^{1-\theta(\gamma)\alpha/(\alpha+1)-\varepsilon\alpha/(\alpha+1)}$$

Since the above inequality holds for any T sufficiently large, this implies that for all $\varepsilon > 0$

$$\theta(\alpha) + \varepsilon \geqslant 1 - \theta(\gamma)\frac{\alpha}{\alpha+1} - \varepsilon\frac{\alpha}{\alpha+1},$$

which yields the desired result as $\varepsilon \to 0$. □

3.2.3. Yet can we adapt in some way?

We have described in (3.7) the minimal rate functions that are compatible with the lower bounds of adaptation: no algorithm can enjoy uniformly better rates. Of course, at this point, the next natural question is whether any of these adaptive rate functions may indeed be reached by an algorithm.

All previous algorithms for continuum-armed bandits require the regularity as an input in some way (see the literature discussion in Section 3.1.1). Such algorithms are flawed: if the true regularity is underestimated then we only recover the guarantees that correspond to the smaller regularity, which is often far worse than the lower bounds of Theorem 3.1. More dramatically, if the true regularity is overestimated, then, a priori, no guarantees hold at all.

We prove that all these rate functions may be achieved by a new algorithm. More precisely, if the player wishes to reach one of the lower bounds θ_m, she may select a value of the input accordingly and match the chosen θ_m. This is our main contribution and is described in the next section.

3.3. An admissible adaptive algorithm and its analysis

We discuss in Section 3.3.1 how the well-known CAB1 algorithm can be generalized for our purpose. In Section 3.3.2 we describe our algorithm and the main upper bound on its regret. Section 3.3.4 is devoted to the anytime version of the algorithm and to a discussion on optimality.

3.3.1. An abstract version of CAB1 as a building block towards adaptation

We describe a generalization of the CAB1 algorithm from Kleinberg [2004], where we include arbitrary measures in the discretization. Although this extension is straightforward, we detail it as we will use this algorithm repeatedly further in this chapter. In the original CAB1, the space of arms is discretized into a partition of K subsets, and an algorithm for finite-armed bandits plays on the K midpoints of the sets. Auer et al. [2007] replace the midpoints by a random point uniformly chosen in the subset.

We introduce a generic version of this algorithm we call AbCAB, for Abstract Continuum-Armed Bandits). We consider K arbitrary probability distributions over $\mathcal{X}$, which we denote by $(\pi_i)_{1\leqslant i\leqslant K}$. Denote also by $\pi(f)$ the expectation of $f(X)$ when $X \sim \pi$. At each time step, the decision maker chooses one distribution, π_{I_t}, and plays an arm picked according to that distribution. By the tower rule, she receives a reward such that

$$\mathbb{E}[Y_t \mid I_t] = \mathbb{E}[f(X_t) \mid I_t] = \pi_{I_t}(f)\,.$$

As the player uses a finite-arm algorithm $\mathcal{A}$ to select I_t, the regret she suffers can be decomposed as the sum of two terms (denoting by $\tilde{R}_T$ the expected regret of the finite-armed algorithm):

$$\overline{R}_T = T\big(M(f) - \max_{i=1,\dots,K} \pi_i(f)\big) + \tilde{R}_T\big(\big(\pi_i(f)\big)_{1\leqslant i\leqslant K}; \mathcal{A}\big)\,. \tag{3.10}$$

This identity is central to the construction of our algorithm. Using terminology from Auer et al. [2007], the first term measures an *approximation error* of the maximum of f, and the other the

actual *cost of learning* in the approximate problem. Parameters are chosen to balance these two sources of error.

Algorithm 3.1 AbCAB (Abstract Continuum-Armed Bandit, adapted from Kleinberg [2004])

Input: T the time horizon, K probability measures over $\mathcal{X}$ denoted by $\pi_1, \dots, \pi_K$, discrete K-armed bandit algorithm $\mathcal{A}$
for $t = 1, 2, \dots, T$ **do**
  Define I_t the arm in $\{1, \dots, K\}$ recommended by $\mathcal{A}$
  Play $X_t \in \mathcal{X}$ drawn according to π_{I_t}, and receive Y_t such that $\mathbb{E}[Y_t | X_t] = f(X_t)$
  Give Y_t as input to $\mathcal{A}$ corresponding to I_t
end for

The canonical example is that for which the space of arms is cut into a partition. Denote by $\mathbf{Disc}(K)$ the family of the uniform measures over the intervals $[(i-1)/K, i/K]$ for $1 \leqslant i \leqslant K$. We state and prove (for completeness) this result to recall the non-adaptive minimax bound (3.1).

Proposition 3.2. *Let $\alpha > 0$ and $L > 1/\sqrt{T}$ be regularity parameters, and define the number of discrete arms $K^\star = \min\left(\left\lceil L^{2/(2\alpha+1)} T^{1/(2\alpha+1)} \right\rceil, T\right)$. Algorithm AbCAB run with the uniform discretization $\boldsymbol{Disc}(K^\star)$ and $\mathcal{A}$ =MOSS enjoys the bound*

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 28\, L^{1/(2\alpha+1)}\, T^{(\alpha+1)/(2\alpha+1)}\,. \tag{3.11}$$

Proof. Choose $f \in \mathcal{H}(L, \alpha)$. Let us denote by $i^\star$ an integer such that there exists an optimal arm $x^\star$ in the interval $\left[(i^\star - 1)/K^\star, i^\star/K^\star\right]$. By the Hölder assumption

$$\frac{1}{K^\star} \int_{(i^\star-1)/K^\star}^{i^\star/K^\star} \big(f(x^\star) - f(x)\big)\, \mathrm{d}x \leqslant L \left(\frac{1}{K^\star}\right)^\alpha ,$$

and this upper bounds the approximation error of the discretization. Moreover, since $T \geqslant K^\star$, the cost of learning is smaller than $18\sqrt{K^\star T}$. Thus by (3.10)

$$\overline{R}_T \leqslant TL \left(\frac{1}{K^\star}\right)^\alpha + 18\sqrt{K^\star T}\,.$$

$K^\star$ was chosen to minimize this quantity. We distinguish cases depending on the value of $K^\star$.

If $1 < K^\star < T$, then $L^{2/(2\alpha+1)} T^{1/(2\alpha+1)} \leqslant K^\star \leqslant 2L^{2/(2\alpha+1)} T^{1/(2\alpha+1)}$ (the bound $\lceil x \rceil \leqslant 2x$, which is valid when $x \geqslant 1$, is more practical to handle the multiplicative constants), we deduce the upper bound:

$$\left(1 + 18\sqrt{2}\right) L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)}\,.$$

Since we assumed that $L > 1/\sqrt{T}$, we have always $K^\star > 1$. Therefore the last case to consider is if $K^\star = T$. Then $L^{2/(2\alpha+1)} T^{1/(2\alpha+1)} \geqslant T/2$ and thus $L \geqslant 2^{-(2\alpha+1)/2}\, T^\alpha$. In this case $L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)} \geqslant (\sqrt{2}/2)T$ and the claimed bound is met since in that case, we have by a trivial bound $\overline{R}_T \leqslant T \leqslant \sqrt{2}\, L^{1/(2\alpha+1)} T^{(\alpha+1)/(2\alpha+1)}$. □

3.3.2. Memorize past plays, Discretize the arm space, and Zoom Out: the MeDZO algorithm

To achieve adaptation, we combine two tricks: *going from fine to coarser discretizations* while *keeping a summary of past plays in memory.*

Our algorithm works in successive regimes. At each time regime i, we reset the algorithm and start over a new regime of length double the previous one ($\Delta T_i = 2^{p+i}$), and with fewer discrete

arms ($K_i = 2^{p+2-i}$). While doing this, we keep in memory the previous plays: in addition to the uniform distributions over the subsets of partitions, we include the empirical measures $\widehat{\nu}_j$ of the actions played in the past regimes, for $j < i$.

Algorithm 3.2 MeDZO (Memorize, Discretize, Zoom Out)

1: **Input**: parameter B, time horizon T
2: **Set**: $p = \lceil \ln_2 B \rceil$, $K_i = 2^{p+2-i}$ and $\Delta T_i = 2^{p+i}$
3: **for** $i = 1, \dots, p$ **do**
4: For ΔT_i rounds, run algorithm AbCAB with the uniform discretization in K_i pieces *and* the empirical measures of the previous plays $\widehat{\nu}_j$ for $j < i$; use MOSS as the discrete algorithm.[a]
5: **Set**: $\widehat{\nu}_i$ the empirical measure of the plays during regime i.
6: **end for**

[a] No $\widehat{\nu}$ is used for $i = 0$

Our construction is based on the following remark. Consider the approximation error suffered during regime i. Denoting the by Π_i the set of measures given to the player during regime i, that is, the uniform measures over the regular K_i-partition and the empirical measures of arms played during the regimes $j < i$, the approximation error is bounded as follows:

$$\Delta T_i \left(M(f) - \mathbb{E}\Big[\max_{\pi \in \Pi_i} \pi(f)\Big] \right) \leqslant \Delta T_i \left(M(f) - \mathbb{E}[\widehat{\nu}_j(f)] \right) = \frac{\Delta T_i}{\Delta T_j} \sum_{t \in \text{Regime } j} \left(M(f) - \mathbb{E}[f(X_t)] \right) \quad (3.12)$$

and this bound is proportional to the regret suffered during regime j. This means that even though we zoom out by using fewer arms, we can make sure that the average approximation error in regime i is less than the regret previously suffered. Moreover, the first discretizations are fine enough to ensure a small regret in the first regimes, thanks to the Hölder property. This argument is formalized in the proof (Lemma 3.1), and shows that MeDZO maintains a balance between approximation and cost of learning that yields optimal regret.

A surprising fact here is that we go from finer to coarser discretizations during the different phases. Thus, paradoxically, *the algorithm zooms out as time passes.* Note also that although this regime-based approach is reminiscent of the doubling trick, there is an essential difference in that information is carried between the regimes via the distribution of the previous plays.

We first state our central result, a generic bound that holds for any input parameter B. We discuss the optimality of these adaptive bounds in the next subsection.

Theorem 3.2. *Algorithm 3.2 run with the knowledge of T and input $B \geqslant \sqrt{T}$ enjoys the following guarantee: for all $\alpha > 0$ and $L > 0$,*

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 412 \, (\ln_2 B)^{3/2} \max\left(B, \, T L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)} \right). \quad (3.13)$$

We provide some illustrative numerical experiments in Appendix 3.B, comparing the results of MeDZO with other non-adaptive algorithms.

3.3.3. Illustration

We provide a figure to illustrate the behavior of MeDZO in a schematic example, when $B = \sqrt{T}$.

MeDZO starts by playing on a fine discretization with a size of order $\sqrt{T}$, but for a short length of time, of order $\sqrt{T}$. At the end of the first epoch, it memorizes the empirical distribution of the arms played; then it runs a new instance of AbCAB with both the coarser discretization, and the memorized action. This process is repeated until the time horizon is reached.

The payoffs of the memorized actions increase until the size of the discretization reaches a critical value; after that they fluctuate. Therefore MeDZO manages to maintain a regret of order the approximation error at this critical discretization, multiplied by T.

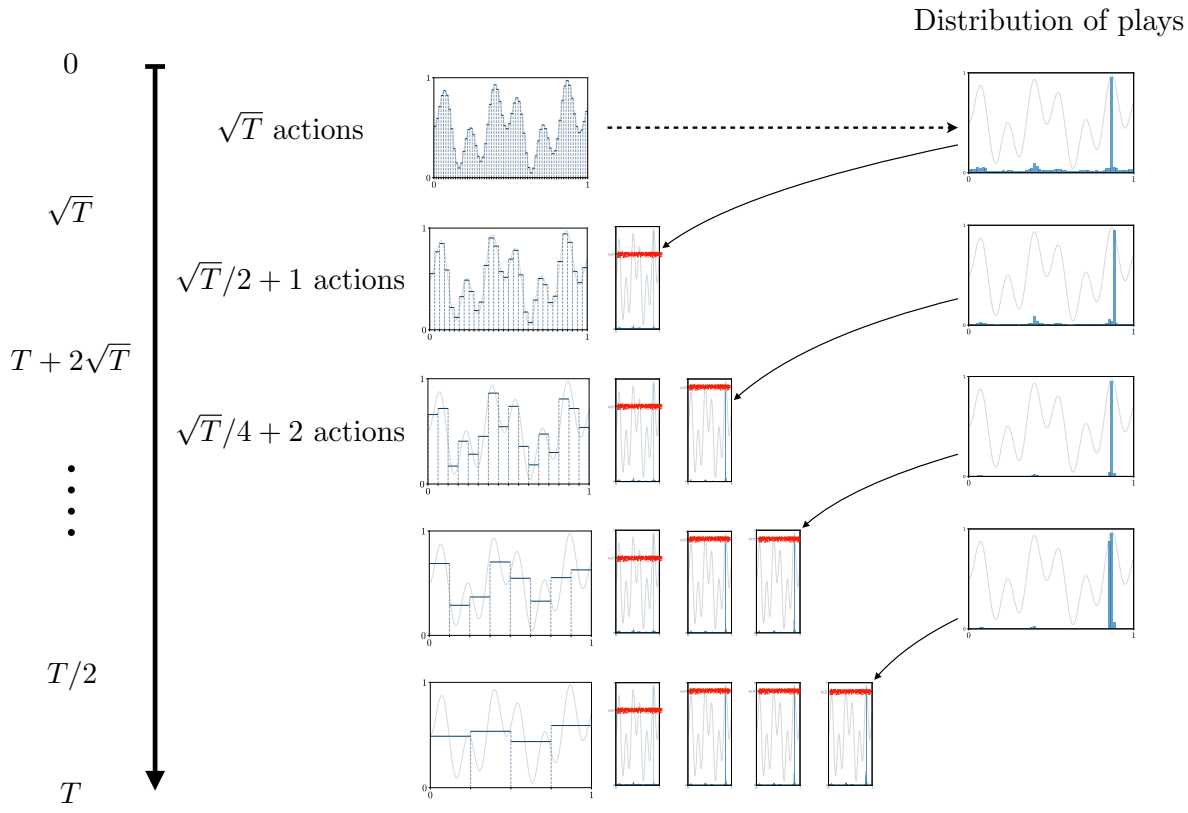


Figure 3.2.: Behavior of MeDZO on a schematic drawing with input $B = \sqrt{T}$. The expected payoffs of the memorized actions are displayed in red; those from the usual discretization are in blue.

3.3.4. Discussion: anytime version and admissibility

Anytime version via the doubling trick The dependence of Algorithm 3.2 on the parameter B makes it horizon-dependent. We use the doubling trick to build an anytime version of the algorithm. At each new doubling-trick regime, we input a value of B that depends on the length of the k-th regime. If it is of length $T^{(k)}$, one typically thinks of $B_k = (T^{(k)})^m$ for some exponent m. In that case, we get the following bound —see the proof and description of the algorithm in Appendix 3.A.

Corollary 3.1 (Doubling trick version). *Choose $m \in [1/2, 1]$. The doubling-trick version of MeDZO, run with m as sole input (and without the knowledge of T) ensures that for all regularity parameters $\alpha > 0$ and $L > 0$ and for $T \geqslant 1$*

$$\sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 4000(\ln_2 T^m)^{3/2} \max\left(T^m, TL^{1/(\alpha+1)}(T^m)^{-\alpha/(\alpha+1)}\right) = \mathcal{O}\big((\ln T)^{3/2}\, T^{\theta_m(\alpha)}\big).$$

Admissibility of Algorithm 3.2 The next result is a direct consequence of Corollary 3.1. This echoes the discussion following Theorem 3.1, and shows that for any input parameter m, the anytime version of MeDZO cannot be improved uniformly for all α.

Corollary 3.2. *For any $m \in [1/2, 1]$, the doubling trick version of MeDZO (see App. 3.A) with input m achieves rate function θ_m, and is therefore admissible.*

3.3.5. About the remaining parameter: the $B = \sqrt{T}$ case

Tuning the value of B amounts to selecting one of the minimal curves in Figure 3.1. Therefore this parameter is a feature of the algorithm, as it allows the player to choose between the possible optimal behaviors. The tuning of this parameter is an unavoidable choice for the player to make.

The next example illustrates well the performance of MeDZO, as it is easily comparable to the usual minimax bounds. Looking at Figure 3.1, this choice corresponds to $m = 1/2$, i.e., the only choice of parameter that reaches the usual minimax rates as $\alpha \to \infty$. In other words, if the players wishes to ensure that her regret on very regular functions is of order $\sqrt{T}$, then she has to pay a polynomial cost of adaptation for not knowing α and that price is exactly the ratio between (3.1) and (3.2).

Corollary 3.3. *Set a horizon T and run Algorithm 3.2 with $B = \sqrt{T}$. Then for $\alpha > 0$ and $L \geqslant 1/\sqrt{T}$,*

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 146\,(\ln_2 T)^{3/2}\, L^{1/(\alpha+1)}\, T^{(\alpha+2)/(2\alpha+2)}\,. \tag{3.14}$$

This is straightforward from Theorem 3.2, since the inequality $B = \sqrt{T} \leqslant T L^{1/(\alpha+1)} \sqrt{T}^{-\alpha/(\alpha+1)}$ holds whenever $L \geqslant 1/\sqrt{T}$. An anytime version of this result can be obtained from Corollary 3.1.

3.4. Proof of Theorem 3.2

Full proof of Theorem 3.2. Let $\mathcal{F}_t = \sigma(I_1, X_1, Y_1, \dots, I_t, X_t, Y_t)$ be the σ-algebra corresponding to the information available at the end of round t. Define also the transition times $T_i = \sum_{j=1}^{i} \Delta T_j$ with the convention $T_0 = 0$. Let us first verify that T is smaller than the total length of the regimes. By definition of p, we have $B \leqslant 2^p < 2B$. Thus $T_p = 2^{p+1}(2^p - 1) \geqslant 2B(B-1) > B^2 > T$, and the algorithm is indeed well-defined up to time T.

Consider the regret suffered during the i-th regime $\overline{R}_{T_{i-1},T_i} := \Delta T_i\, M(f) - \sum_{t=T_{i-1}+1}^{T_i} \mathbb{E}\big[f(X_t)\big]$. We bound this quantity thanks to the decomposition (3.10), by first conditioning on the past up to time T_{i-1}. Since there are $K_i + i$ discrete actions, the regret bound on MOSS ensures that

$$\mathbb{E}\left[\sum_{t=T_{i-1}+1}^{T_i} \big(M(f) - f(X_t)\big) \,\middle|\, \mathcal{F}_{T_{i-1}}\right] \leqslant \Delta T_i \big(M(f) - M_i^\star\big) + 18\sqrt{(K_i + i)\Delta T_i} \tag{3.15}$$

where $M_i^\star = \max\{\pi_j^{(i)}(f) \mid \pi_j^{(i)} \in \mathbf{Disc}(K_i)\} \cup \big\{\widehat{\nu}_\ell(f) \mid \ell = 0, \dots, i-1\big\}$. Notice that this bound holds even though $M_i^\star$ is a random variable, as the algorithm is completely reset, and the measures $(\widehat{\nu}_j)_{j<i}$ are fixed at time $T_{i-1} + 1$ (i.e., they are $\mathcal{F}_{T_{i-1}}$-measurable). Integrating once more, we obtain

$$\overline{R}_{T_{i-1},T_i} \leqslant \Delta T_i \big(M(f) - \mathbb{E}[M_i^\star]\big) + 18\sqrt{(K_i + i)\Delta T_i}\,. \tag{3.16}$$

Bounding the cost of learning. By definition of K_i and ΔT_i, we have $K_i \Delta T_i = 2^{2p+2} \leqslant 16 B^2$. Therefore, since p and K_i are integers greater than 1, using $a + b - 1 \leqslant ab$ for positive integers,

$$\sqrt{(K_i + i)\Delta T_i} \leqslant \sqrt{(K_i + p - 1)\Delta T_i} \leqslant \sqrt{p K_i \Delta T_i} \leqslant 4\sqrt{p} B\,. \tag{3.17}$$

Bounding the approximation error. The key ingredient for this part is the following fact, that synthetizes the benefits of our construction as hinted in (3.12) and the surrounding discussion.

Lemma 3.1. *The total approximation error of MeDZO in regime i is controlled by the Hölder bound on the grid of mesh size $1/K_i$, and by the regret suffered during the previous regimes,*

$$\Delta T_i \left(M(f) - \mathbb{E}[M_i^\star]\right) \leqslant \Delta T_i \, \min\left(L \, \frac{1}{K_i^\alpha}, \min_{j<i} \left(\frac{\overline{R}_{T_{j-1},T_j}}{\Delta T_j} \right) \right) \tag{3.18}$$

Proof. This derives easily from the construction of the algorithm, i.e., from the definition of $M_i^\star$. Considering an interval in the regular K_i-partition that contains a maximum of f, by the Hölder property, $M(f) - M_i^\star \leqslant L/K_i^\alpha$. For the second minimum, as described in Eq. (3.12), for $j < i$,

$$M(f) - M_i^\star \leqslant M(f) - \widehat{\nu}_j(f) = \frac{1}{\Delta T_j} \sum_{t=T_{j-1}+1}^{T_j} \big(M(f) - f(X_t)\big) \, .$$

Taking an expectation, $\overline{R}_{T_{j-1},T_j}$ appears, and we conclude by taking the minimum over j. □

Remember that since $K_i \Delta T_i = 2^{2p+2}$, we have $L \, \Delta T_i / K_i^\alpha = L 2^{2p+2} / K_i^{1+\alpha}$. Therefore, the first bound on the approximation error in (3.18) increases with i, as K_i decreases with i. Denote by i_0 the last time regime i for which

$$L \frac{\Delta T_{i_0}}{K_{i_0}^\alpha} \leqslant B \, . \tag{3.19}$$

If this is never satisfied, i.e., not even for $i = 1$, then $L 2^{p+1} / 2^{\alpha(p+1)} > B$ which yields, using $B \leqslant 2^p \leqslant 2B$, that $4LB \geqslant 2^{\alpha+1} B^\alpha B$ and then $L > B^\alpha / 2$. In that case, $L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)} \geqslant 1$ and the total regret bound (3.13) is true as it is weaker than the trivial bound $R_T \leqslant T$.

Hence we may assume that $i_0 \geqslant 1$ is well defined. By comparing i to i_0, we now show the inequality

$$\sum_{i=1}^{p} \Delta T_i \big(M(f) - \mathbb{E}[M_i^\star]\big) \leqslant \sum_{i=1}^{i_0} B + \sum_{i=i_0+1}^{p} 2(1 + 72\sqrt{p}) \Delta T_i \, L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)} \, . \tag{3.20}$$

For all $i \leqslant i_0$ the approximation error is smaller than the first argument of the minimum in (3.18), and this term is smaller than B. Therefore $\Delta T_i \big(M(f) - \mathbb{E}[M_i^\star]\big) \leqslant B$. In particular, this together with (3.16) and (3.17) implies that the total regret suffered during regime i_0 is $\overline{R}_{T_{i_0-1},T_{i_0}} \leqslant (1 + 72\sqrt{p}) B$.

For the later time regimes $i_0 < i \leqslant p$, we use the fact that preceding empirical measures were kept as discrete actions, and in particular the one of the i_0-th regime: (3.18) instantiated with $j = i_0$ yields

$$\Delta T_i \left(M(f) - \mathbb{E}[M_i^\star]\right) \leqslant \Delta T_i \frac{\overline{R}_{T_{i_0-1},T_{i_0}}}{\Delta T_{i_0}} \leqslant \big(1 + 72\sqrt{p}\big) \Delta T_i \frac{B}{\Delta T_{i_0}} \, . \tag{3.21}$$

Solving equations $L \Delta T_{i_0} / K_{i_0}^\alpha \approx B \approx 4\sqrt{\Delta T_{i_0} K_{i_0}}$, we get $B / \Delta T_{i_0} \leqslant 2 \, L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)}$, (details are given after the proof). Therefore for $i_0 < i \leqslant p$, using (3.21),

$$\Delta T_i \big(M(f) - \mathbb{E}[M_i^\star]\big) \leqslant 2(1 + 72\sqrt{p}) \, \Delta T_i \; L^{1/(\alpha+1)} B^{-\alpha/(\alpha+1)} \, ,$$

and we obtain (3.20) by summing over i.

Conclusion We conclude with some crude boundings. First, as $i_0 \leqslant p$ and the sum of the ΔT_i's is smaller than T, the total approximation error is less than $pB+2(1+72\sqrt{p})TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}$. Let us include the cost of learning, which is smaller than $72p\sqrt{p}B$ and conclude, using $a+b \leqslant \max(a,b)$

$$\begin{aligned}\overline{R}_T &\leqslant 2(1+72\sqrt{p})TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)} + pB + 72p^{3/2}B \\ &= 2(1+72\sqrt{p})TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)} + p(1+72\sqrt{p})B \\ &\leqslant \Big(2(1+72\sqrt{p}) + p\big(1+72\sqrt{p}\big)\Big)\max\big(B,\, TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}\big)\end{aligned} \tag{3.22}$$

from which the desired bound follows, using $1 \leqslant p$, and $p \leqslant 2\ln_2 B$ and $4(1+72\sqrt{2}) \leqslant 412$. □

Details on (3.20)*, in the proof of Theorem 3.2.* By definition of i_0, and since we assumed that $i_0 < p$

$$B \leqslant L\,\frac{\Delta T_{i_0+1}}{K_{i_0+1}^{\alpha}}\,,$$

i.e., using $K_{i_0}\,\Delta T_{i_0} = 2^{2p+2}$,

$$B \leqslant 2^{1+\alpha}L\,\frac{\Delta T_{i_0}}{K_{i_0}^{\alpha}} = 2^{1+\alpha}L\,(\Delta T_{i_0})^{1+\alpha}\,2^{-(2p+2)\alpha}\,.$$

From this we deduce, using $2^p \geqslant B$ for the second inequality,

$$\big(\Delta T_{i_0}\big)^{(1+\alpha)} \geqslant 2^{-1-\alpha}BL^{-1}2^{(2p+2)\alpha} \geqslant 2^{-1+\alpha}L^{-1}B^{2\alpha+1}\,.$$

Hence, using $2^{(\alpha-1)/(\alpha+1)} \geqslant 1/2$, we obtain $\Delta T_{i_0} \geqslant (1/2)L^{-1/(\alpha+1)}B^{(2\alpha+1)/(\alpha+1)}$, thus $B/\Delta T_{i_0} \leqslant 2\,L^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}$. □

3.5. Further considerations

Local regularity assumption Theorem 3.2 holds under a relaxed smoothness assumption, namely that the function satisfies the Hölder condition only in a small cell containing the maximum. By looking carefully at the proof, we observe that the condition is only required up to the i_0-th epoch (defined in (3.19)), at which the size of the cells in the discretization is of order $1/K_{i_0} \approx (LB)^{-1/(1+\alpha)}$. Therefore we only need condition (3.4) to be satisfied for points x in an interval of size $(LB)^{-1/(1+\alpha)}$ around the maximum.

Higher dimension Our results can be generalized to functions $[0,1]^d \to [0,1]$ that are $\|\cdot\|_\infty$-Hölder. For MeDZO to be well-defined, take $K_i = 2^{d(p+2-i)}$ and $\Delta T_i = 2^{d(p+i)}$, with $p \approx (\ln B)/d$. The bounds are similar to their one-dimensional counterparts, up to replacing α by α/d in the exponents, but the constants are deteriorated by a factor that is exponential in d. The bound in Theorem 3.2 changes to $\max\big(B, L^{d/(\alpha+d)}TB^{-\alpha/(\alpha+d)}\big)$.

Appendix for Chapter 3

3.A. Anytime-MeDZO and its analysis

The doubling trick is the most standard way of converting non-anytime algorithms into anytime algorithms, when the regret bound is polynomial. It consists in taking fresh starts of the algorithm over a grid of dyadic times. The implementation of the trick is straightforward in our case.

Algorithm 3.3 Doubling trick MeDZO

1: **Input**: parameter $m \in [1/2, 1]$;
2: **for** $i = 0, \dots$ **do**
3: Run MeDZO (Alg. 3.2) with input $B = 2^{im}$ for 2^i rounds
4: **end for**

Corollary (Doubling trick version). *Choose $m \in [1/2, 1]$. The doubling-trick version of MeDZO, run with m as sole input (and without the knowledge of T) ensures that for all regularity parameters $\alpha > 0$ and $L > 0$ and for $T \geqslant 1$*

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{R}_T \leqslant 4000 (\ln_2 T^m)^{3/2} \max\left(T^m, T L^{1/(\alpha+1)} (T^m)^{-\alpha/(\alpha+1)}\right) = \mathcal{O}\left((\ln T)^{3/2} T^{\theta_m(\alpha)}\right).$$

As the regret bound is not exactly of the form cT^θ, we work with the polynomial version of the bound on the regret of MeDZO, equation (3.22), for the doubling trick to be effective. Obviously the value of the constant in the bound is not our main focus, but we still write it explicitly as it shows that there is no hidden dependence on the various parameters.

Proof. By (3.22), with $p_i = \lceil \ln_2 2^{im} \rceil \leqslant 1 + \ln_2 2^{im}$, in the i-th doubling trick regime, the cumulative regret is bounded by

$$2(1 + 72\sqrt{1 + \ln_2 2^{im}}) 2^i L^{1/(\alpha+1)} (2^{im})^{-\alpha/(\alpha+1)} + (1 + \ln_2 2^{im})\left(1 + 72\sqrt{1 + \ln_2 2^{im}}\right) 2^{im}$$

Now since

$$\sum_{i=0}^{\lceil \ln_2 T \rceil} 2^i = 2^{\lceil \ln_2 T \rceil + 1} - 1 \geqslant 2T - 1 \geqslant T\,,$$

there are always less than $\lceil \ln_2 T \rceil$ full regimes. Therefore, using $\ln_2 2^{im} \leqslant \ln_2 T^m$, and summing over the regimes, the first part of this sum is bounded by

$$\begin{aligned}
&2(1 + 72\sqrt{2 \ln_2 T^m}) L^{1/(\alpha+1)} \sum_{i=0}^{\lceil \ln_2 T \rceil} 2^{i(1 - m\alpha/(\alpha+1))} \\
&\leqslant 2(1 + 72\sqrt{2 \ln_2 T^m}) L^{1/(\alpha+1)} \frac{2^{(\lceil \ln_2 T \rceil + 1)(1 - m\alpha/(\alpha+1))}}{2^{1 - m\alpha/(\alpha+1)} - 1} \\
&\leqslant 2(1 + 72\sqrt{2}) \sqrt{\ln_2 T^m} L^{1/(\alpha+1)} \frac{2^{2(1 - m\alpha/(\alpha+1))}}{\sqrt{2} - 1} T (T^m)^{-\alpha/(\alpha+1)} \\
&\leqslant 2(1 + 72\sqrt{2}) \sqrt{\ln_2 T^m} L^{1/(\alpha+1)} \frac{4}{\sqrt{2} - 1} T (T^m)^{-\alpha/(\alpha+1)}
\end{aligned}$$

where we used $2^{\lceil \ln_2 T \rceil} \leqslant 2T$; we also used the fact that since $m \geqslant 1/2$, we always have the inequality $1 - m\alpha/(\alpha+1) \geqslant 1/2$ to bound the denominator. Similarly, the second part is bounded

by

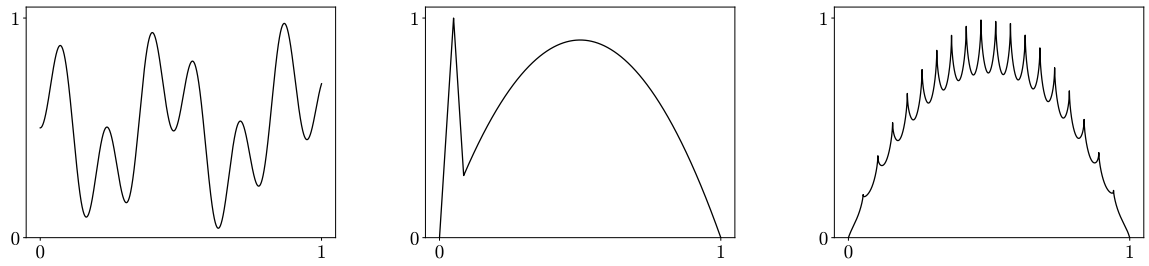
$$2(1+72\sqrt{2})(\ln_2 T^m)^{3/2}\sum_{i=0}^{\lceil \ln_2 T\rceil} 2^{im} \leqslant 2(1+72\sqrt{2})(\ln_2 T^m)^{3/2}\frac{4}{\sqrt{2}-1}T^m\,.$$

All in all, we obtain the same minimax guarantees as if we had known the time horizon in advance, but with an extra multiplicative factor of $4/(\sqrt{2}-1) \approx 9,66$. □

3.B. Numerical experiments

This section contains some numerical experiments comparing the regrets of algorithms that require the knowledge of the smoothness, against MeDZO.

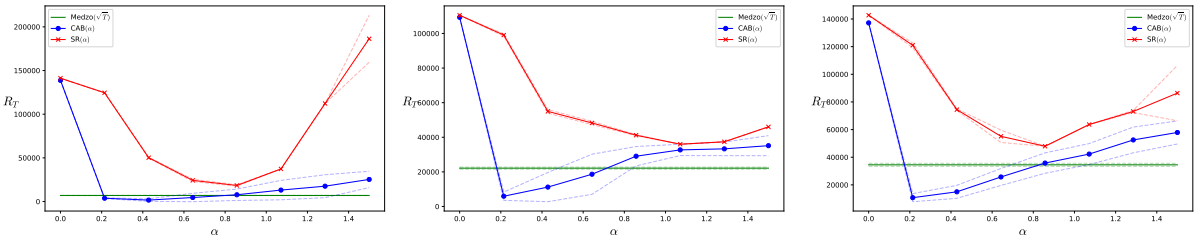
We examine bandit problems defined by their mean-payoff functions and gaussian $\mathcal{N}(0;1/4)$ noise. The functions considered are $f : x \mapsto (1/2)\sin(13x)\sin(27x) + 0.5$ taken from Bubeck et al. [2011b], $g : x \mapsto \max\left(3.6\,x(1-x), 1-1/0.05\,|x-0.05|\right)$ adapted from Coquelin and Munos [2007] and the Garland function $x \mapsto x(1-x)(4-\sqrt{|\sin(60x)|}$, which we took from Valko et al. [2013]. The functions are plotted in Figure 3.3.



(a) f (b) g (c) The Garland function

Figure 3.3.: Problems considered

The algorithms we compare are SR from Locatelli and Carpentier [2018], and CAB1 from Kleinberg [2004] with MOSS as the discrete algorithm. SR takes directly the smoothness α as an input, and assumes $L = 1$. For CAB1, we compute the optimal discretization size for $L = 1$ and varying α.



(a) f (b) g (c) The Garland function

Figure 3.4.: Regrets of MeDZO, and of SR and CAB1 run with different values of the smoothness parameter.

In Figure 3.4 we plot the cumulative regret of the algorithms after a time horizon $T = 300000$, for varying values of the assumed smoothness. For each problem, MeDZO was run only once, as it does not need to know the smoothness. The regret was averaged over $N = 75$ runs, and the dotted curves represent +/- one standard deviation.

We recall that minimax guarantees are worst-case guarantees, therefore comparing algorithms on a single problem can only serve as an empirical illustration.

As expected, the regrets of both SR and CAB1 depend on some careful tuning of the input parameter, determined by the smoothness. The optimal tuning is unclear, and seems to vary on the algorithm. MeDZO, on the other hand, obtains reasonable regret with no tuning. Surprisingly, CAB1 with overestimated smoothness seems to behave quite well, although the large variance sometimes makes it difficult to distinguish the results. Recall that MeDZO is the only algorithm with theoretical guarantees for high values of α.

3.C. About simple regret

In this section, we consider the case of *simple regret*, which complements the discussion about adaptation to smoothness in sequential optimization procedures. We write out how to achieve adaptation at usual rates for simple regret under Hölder smoothness assumptions. We do not claim novelty here, as adaptive strategies have already been used for simple regret under more sophisticated regularity conditions (see, e.g., Grill et al. [2015], Shang et al. [2019] and a sketched out procedure in Locatelli and Carpentier [2018]); however, we feel the details deserve to be written out in this simpler setting.

Let us recall the definition of simple regret. In some cases, we may only require that the algorithm outputs a recommendation $\tilde{X}_T$ at the end of the T rounds, with the aim of minimizing the simple regret, defined as

$$\overline{r}_T = M(f) - \mathbb{E}\big[f\big(\tilde{X}_T\big)\ \big]\,.$$

This setting is known under various names, e.g., pure exploration, global optimization or black-box optimization. As noted in Bubeck et al. [2011a], minimizing the simple regret is easier than minimizing the cumulative regret in the sense that if the decision-maker chooses a recommendation uniformly among the arms played $X_1, \dots, X_T$, then

$$\overline{r}_T = M(f) - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[f\big(X_t\big)\big] = \frac{\overline{R}_T}{T}\,. \tag{3.23}$$

The minimax rates of simple regret over Hölder classes $\mathcal{H}(L,\alpha)$ are lower bounded by $\Omega(L^{1/(2\alpha+1)}T^{-\alpha/(2\alpha+1)})$, which are exactly the rates for cumulative regret divided by T (see Locatelli and Carpentier [2018] for a proof of the lower bound). Consequently, at known regularity, any minimax optimal algorithm for cumulative regret automatically yields a minimax recommendation for simple regret via (3.23).

When the smoothness is unknown, the situation turns out to be quite different. Adapting to the Hölder parameters can be done at only a (poly-)logarithmic cost for simple regret, contrasting with the polynomial cost of adaptation of cumulative regret. This can be achieved thanks to a very general and simple cross-validation scheme defined in Shang et al. [2019], named General Parallel Optimization.

Algorithm 3.4 GPO (General Parallel Optimization) for Hölder minimax adaptation

Input: time horizon $T \geqslant 8$
Set: $p = \lceil \ln_2 T \rceil$ and define $K_i = 2^i$ for $i = 1, \dots, p$
for $i = 1, \dots, p$ **do** // Exploration
  For $\lfloor T/(2p) \rfloor$ rounds, run algorithm AbCAB with the discretization in K_i pieces; use MOSS as the discrete algorithm
  Define output recommendation $\tilde{X}^{(i)}$, uniformly chosen among the $\lfloor T/(2p) \rfloor$ arms played
end for
for $i = 1, \dots, p$ **do** // Cross-validation
  Play $\lfloor T/(2p) \rfloor$ times each recommendation $\tilde{X}^{(i)}$ and compute the average reward $\widehat{\mu}^{(i)}$
end for
return A recommendation $\tilde{X}_T = \tilde{X}^{(\widehat{\imath})}$ with $\widehat{\imath} \in \operatorname{argmax} \widehat{\mu}^{(i)}$

The next result shows that the player obtains the same simple regret bounds as when the smoothness is known (up to logarithmic factors).

Theorem 3.3. *GPO with AbCAB as a sub-algorithm (Alg. 3.4) achieves, given $T \geqslant 8$ and without the knowledge of α and L, for all $\alpha > 0$ and $L \geqslant 2^{\alpha+1/2}\sqrt{\lceil \ln_2 T \rceil / T}$ the bound*

$$\sup_{f \in \mathcal{H}(L,\alpha)} \overline{r}_T \leqslant \Big(54 + \frac{\sqrt{\pi}}{2} \ln_2 T\Big) L^{1/(2\alpha+1)} \left(\frac{\lceil \ln_2 T \rceil}{T} \right)^{\alpha/(2\alpha+1)} = \tilde{\mathcal{O}}\Big(L^{1/(2\alpha+1)} T^{-\alpha/(2\alpha+1)} \Big) \, .$$

The $\tilde{\mathcal{O}}$ notation hides the $\ln T$ factors, and the assumption that $T \geqslant 8$ is needed to ensure that $T/(2p) = T/(2\lceil \ln_2 T \rceil) \geqslant 1$: otherwise the algorithm itself is ill-defined.

Proof. Let $f \in \mathcal{H}(L, \alpha)$ denote a mean-payoff function. Once again we decompose the error of the algorithm into two sources. The simple regret is the sum of the regret of the best recommendation among the p received, $r_{\min}$, and of a cross-validation error, r_{CV},

$$M(f) - \mathbb{E}[f(\tilde{X}_T)] = \underbrace{\min_{i=1,\dots,p} \left(M(f) - \mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big) \Big] \right)}_{r_{\min}} + \underbrace{\max_{i=1,\dots,p} \left(\mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big) \Big] - \mathbb{E}\Big[f\big(\tilde{X}_T\big) \Big] \right)}_{r_{\mathrm{CV}}} . \tag{3.24}$$

We now show that $r_{\mathrm{CV}} \leqslant p^{3/2}\sqrt{\pi/(4T)}$, by detailing an argument that is sketched in the proof of Thm. 3 in Shang et al. [2019]. Denote by $\widehat{\mu}^{(i)}$ the empirical reward associated to recommendation i, and $\widehat{\imath} = \operatorname{argmax} \widehat{\mu}^{(i)}$, so that $\tilde{X}_T = \tilde{X}^{(\widehat{\imath})}$. Then for any fixed i, by the tower rule,

$$\mathbb{E}\big[\widehat{\mu}^{(i)}\big] = \mathbb{E}\Big[\mathbb{E}\Big[\widehat{\mu}^{(i)} \,\Big|\, \tilde{X}^{(i)} \Big] \Big] = \mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big) \Big] \, . \tag{3.25}$$

Therefore, by the above remarks, and since $\widehat{\mu}^{(i)} \leqslant \widehat{\mu}^{(\widehat{\imath})}$,

$$\mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big) \Big] - \mathbb{E}\Big[f\big(\tilde{X}_T\big) \Big] = \mathbb{E}\Big[\widehat{\mu}^{(i)} - f\big(\tilde{X}^{(\widehat{\imath})}\big) \Big] \leqslant \mathbb{E}\Big[\widehat{\mu}^{(\widehat{\imath})} - f\big(\tilde{X}^{(\widehat{\imath})}\big) \Big] \, .$$

We have to be careful here, as $\widehat{\imath}$ is a random index that depends on the random variables $\widehat{\mu}^{(i)}$'s: we cannot apply directly the tower rule as in (3.25). To deal with this, let us use an integrated union bound. Denote by $(\,\cdot\,)^+$ the positive part function, then

$$\mathbb{E}\Big[\widehat{\mu}^{(\widehat{\imath})} - f\big(\tilde{X}^{(\widehat{\imath})}\big) \Big] \leqslant \mathbb{E}\left[\Big(\widehat{\mu}^{(\widehat{\imath})} - f\big(\tilde{X}^{(\widehat{\imath})}\big) \Big)^+ \right] \leqslant \sum_{j=1}^{p} \mathbb{E}\left[\Big(\widehat{\mu}^{(j)} - f\big(\tilde{X}^{(j)}\big) \Big)^+ \right] ,$$

and we are back to handling empirical means of i.i.d. random variables. For each j, the reward given $\tilde{X}^{(j)}$ is $(1/4)$-subgaussian. Therefore, as $\widehat{\mu}^{(i)}$ is the empirical mean of $n = \lfloor T/(2p) \rfloor$ plays of the same arm $\tilde{X}^{(j)}$, this mean $\widehat{\mu}^{(i)}$ is $(1/(4n))$-subgaussian conditionally on $\tilde{X}^{(j)}$ and thus for all $\varepsilon > 0$

$$\mathbb{P}\Big[\widehat{\mu}^{(j)} - f\big(\tilde{X}^{(j)}\big) \geqslant \varepsilon\Big] \leqslant e^{-2n\varepsilon^2}\,.$$

Hence by integrating over $\varepsilon \in [0, +\infty)$, using Fubini's theorem, a change of variable $x = \sqrt{4n}\varepsilon$ (and using the fact that $\lfloor T/(2p) \rfloor \geqslant T/(4p)$ as $T/(2p) \geqslant 1$):

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\mu}^{(j)} - f\big(\tilde{X}^{(j)}\big)\right)^+\right] &= \int_0^{+\infty} \mathbb{P}\Big[\widehat{\mu}^{(j)} - f\big(\tilde{X}^{(j)}\big) \geqslant \varepsilon\Big]\,\mathrm{d}\varepsilon \\
&\leqslant \int_0^{+\infty} e^{-2n\varepsilon^2}\,\mathrm{d}\varepsilon = \frac{1}{\sqrt{4n}} \int_0^{+\infty} e^{-x^2/2}\,\mathrm{d}x \\
&= \sqrt{\frac{\pi}{8n}} = \sqrt{\frac{\pi}{8\,\lfloor T/2p \rfloor}} \leqslant \sqrt{\frac{\pi p}{4\,T}}
\end{aligned}$$

Putting back the pieces together, we have shown that for any i,

$$\mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big)\Big] - \mathbb{E}\Big[f\big(\tilde{X}_T\big)\Big] \leqslant \sum_{j=1}^{p} \sqrt{\frac{\pi p}{4\,T}} = p^{3/2}\sqrt{\frac{\pi}{4\,T}}\,.$$

We deduce the same bound for r_{CV} by taking the maximum over i.

Let us now bound $r_{\min}$. By Eq. (3.10), using the fact that $\lfloor T/(2p) \rfloor \geqslant T/(4p)$ as $T/(2p) \geqslant 1$, for all i

$$M(f) - \mathbb{E}\Big[f\big(\tilde{X}^{(i)}\big)\Big] \leqslant \frac{L}{K_i^{\alpha}} + 18\sqrt{\frac{4pK_i}{T}}\,.$$

We summarize a few calculations in the next lemma. These calculations come from the minimization over the K_i's of the previous bound, with a case disjunction arising from the boundary cases.

Lemma 3.2. *At least one of the three following inequalities holds :*

$$L < 2^{\alpha+1/2}\sqrt{\frac{p}{T}} \quad \textit{or} \quad L \geqslant T^{\alpha}\sqrt{p}$$

or

$$\min_{i=1,\dots,p} \left(\frac{L}{K_i^{\alpha}} + 36\sqrt{\frac{pK_i}{T}}\right) \leqslant 53 L^{1/(2\alpha+1)} \left(\frac{p}{T}\right)^{\alpha/(2\alpha+1)}\,.$$

Let us consider these three cases separately. The first one is forbidden by the assumption that $L \geqslant 2^{\alpha+1/2}\sqrt{p/T}$. In the second case, the function is so irregular that the claimed bound becomes worse than $\overline{r}_T \leqslant 56\, p^{1/2+\alpha/(2\alpha+1)}$, which is weaker than the trivial bound $\overline{r}_T \leqslant 1$.

Finally, in the third case, we may assume that $L \geqslant 2^{\alpha+1/2}\sqrt{p/T} \geqslant \sqrt{p/T}$. Then we have

$$L^{1/(2\alpha+1)} \geqslant \left(\frac{p}{T}\right)^{1/(2(2\alpha+1))} = \left(\frac{p}{T}\right)^{1/2} \left(\frac{p}{T}\right)^{-\alpha/(2\alpha+1)}\,,$$

and thus $\sqrt{p/T} \leqslant L^{1/(2\alpha+1)}(p/T)^{\alpha/(2\alpha+1)}$. By injecting the bound of Lemma 3.2 and the bound on r_{CV} into (3.24):

$$\overline{r}_T \leqslant 53 L^{1/(2\alpha+1)} \left(\frac{p}{T}\right)^{\alpha/(2\alpha+1)} + p\sqrt{\frac{\pi}{4}}\sqrt{\frac{p}{T}} \leqslant (53 + p\sqrt{\pi/4}) L^{1/(2\alpha+1)} \left(\frac{p}{T}\right)^{\alpha/(2\alpha+1)}$$

and the stated bound holds, since $53 + p\sqrt{\pi/4} \leqslant 53 + (\ln_2 T + 1)\sqrt{\pi/4} \leqslant 54 + \sqrt{\pi/4}\ln_2 T$. □

Proof of Lemma 3.2. We upper bound the minimum by comparing the two quantities

$$\frac{L}{K_i^\alpha} \quad \text{v.s.} \quad \sqrt{\frac{pK_i}{T}}\,.$$

As the first term is decreasing with i, and the second term is increasing with i, two extreme cases have te be dealt with. If the first term is always smaller than the second, i.e., even for $i = 1$, then:

$$\frac{L}{2^\alpha} < \sqrt{\frac{p\,2}{T}}\,.$$

This is the first case in the statement of the lemma. Otherwise, the first term might always be greater than the second one, i.e., even for $i = p$ and

$$\frac{L}{2^{\alpha p}} \geqslant \sqrt{\frac{p2^p}{T}}$$

which is equivalent to

$$L^2 \geqslant p\frac{2^{p(2\alpha+1)}}{T}\,,$$

hence, since $2^p \geqslant T$,

$$L^2 \geqslant pT^{2\alpha}$$

which is exactly the second inequality of our statement.

Otherwise, define $i^\star$ to be an index such that

$$\frac{L}{K_{i^\star-1}^\alpha} \geqslant \sqrt{\frac{pK_{i^\star-1}}{T}} \quad \text{and} \quad \frac{L}{K_{i^\star}^\alpha} \leqslant \sqrt{\frac{pK_{i^\star}}{T}} \tag{3.26}$$

By the preceding discussion, $i^\star$ is well defined and $1 < i^\star \leqslant p$. Then by definition of $i^\star$ (the first equation in (3.26))

$$2^{\alpha+1/2}\frac{L}{K_{i^\star}^\alpha} \geqslant \sqrt{\frac{pK_{i^\star}}{T}}\,.$$

Hence, by squaring and regrouping the terms

$$K_{i^\star}^{2\alpha+1} \leqslant 2^{2\alpha+1}L^2\frac{T}{p}$$

thus

$$K_{i^\star} \leqslant 2L^{2/(\alpha+1)}\left(\frac{T}{p}\right)^{1/(2\alpha+1)}$$

and

$$\sqrt{\frac{pK_{i^\star}}{T}} \leqslant \sqrt{2}L^{1/(2\alpha+1)}\left(\frac{p}{T}\right)^{\alpha/(2\alpha+1)}$$

and finally, recalling the second equation in (3.26)

$$\frac{L}{K_{i^\star}^\alpha} + 36\sqrt{\frac{pK_{i^\star}}{T}} \leqslant 37\sqrt{\frac{pK_{i^\star}}{T}} \leqslant 37\sqrt{2}L^{1/(2\alpha+1)}\left(\frac{p}{T}\right)^{\alpha/(2\alpha+1)}.$$

□

3.D. Proof of our version of the lower bound of adaptation

Here we provide the full proof of our version of the lower bound of adaptation stated in Section 3.2.2.

Our statement differs from that of Locatelli and Carpentier [2018] on some aspects. First, and most importantly, we include the dependence on the Lipschitz constants, and we do not consider margin regularity. We also remove a superfluous requirement on B, that $B \leqslant cT^{(\alpha+1)/(2\alpha+1)}$, which was just an artifact of the original proof. Furthermore we believe that the additional condition that $L \leqslant \mathcal{O}(T^{\alpha/2})$ in our version was implicitely used in this original proof. Finally, the value of the constant differs, partly because of the analysis, and partly because we consider $(1/4)$-subgaussian noise instead of 1-subgaussian noise.

We managed to obtain these improvements thanks to a different proof technique. In the original proof, the authors compare the empirical likelihoods of different outcomes and use the Bretagnolle-Huber inequality. We choose to build the lower bound in a slightly different way (see Garivier et al. [2019]): we handle the changes of measure implicitly thanks to Pinsker's inequality (Lemma 3.3). Following Lattimore and Szepesvári [2020], we also chose to be very precise in the definition of the bandit model, in order to make rigorous a few arguments that are often used implicitly in the literature on continuous bandits.

The main argument of the proof, that is, the sets of functions considered, are already present in Locatelli and Carpentier [2018].

Before we start with the proof, let us state a technical tool. Denote by KL the Kullback-Leibler divergence. The next lemma is a generalized version of Pinsker's inequality, tailored to our needs.

Lemma 3.3. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures. For any random variable $Z \in [0,1]$,*

$$|\mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z]| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{2}}$$

Proof. For $z \in [0,1]$, by the classical version of Pinsker's inequality applied to the event $\{Z \geqslant z\}$:

$$|\mathbb{P}[Z \geqslant z] - \mathbb{Q}[Z \geqslant z]| \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{2}}\,.$$

Therefore, by Fubini's theorem and the triangle inequality, and by integrating the preceding inequality:

$$|\mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z]| = \left|\int_0^1 \big(\mathbb{P}[Z \geqslant z] - \mathbb{Q}[Z \geqslant z]\big)\mathrm{d}z\right| \leqslant \int_0^1 |\mathbb{P}[Z \geqslant z] - \mathbb{Q}[Z \geqslant z]|\mathrm{d}z \leqslant \sqrt{\frac{\mathrm{KL}(\mathbb{P},\mathbb{Q})}{2}}$$

□

Proof of the lower bound. For the sake of completeness, we recall in detail the construction of Locatelli and Carpentier [2018], with some minor simplifications that fit our setting. Fix regularity parameters ℓ, L, α and γ satisfying $\ell \leqslant L$ and $\gamma \geqslant \alpha$, so that $\mathcal{H}(\ell,\gamma) \subset \mathcal{H}(L,\alpha)$ (remember the functions are defined on $\mathcal{X} = [0,1]$).

Fix $M \in [1/2, 1]$. Let $K \in \mathbb{N} \setminus \{0\}$ and $\Delta \in \mathbb{R}_+$ be some parameters of the construction whose values will be determined by the analysis. We define furthermore a partition of $[0,1]$ into $K+1$ sets, $H_0 = [1/2, 1]$ and $H_i = [(i-1)/(2K), i/(2K)]$ for $1 \leqslant i \leqslant K$, along with their middle points $x_i \in H_i$. Finally, define the set of hypotheses ϕ_i for $i = 0, \dots, K$ as follows

$$\phi_i(x) = \begin{cases} \max\big(M-\Delta,\, M - \Delta/2 - \ell\,|x - x_0|^\gamma\big) & \text{if } x \in H_0\,, \\ \max\big(M-\Delta,\, M - L\,|x - x_i|^\alpha\big) & \text{if } x \in H_i \text{ and } i \neq 0\,, \\ M - \Delta & \text{otherwise.} \end{cases} \tag{3.27}$$

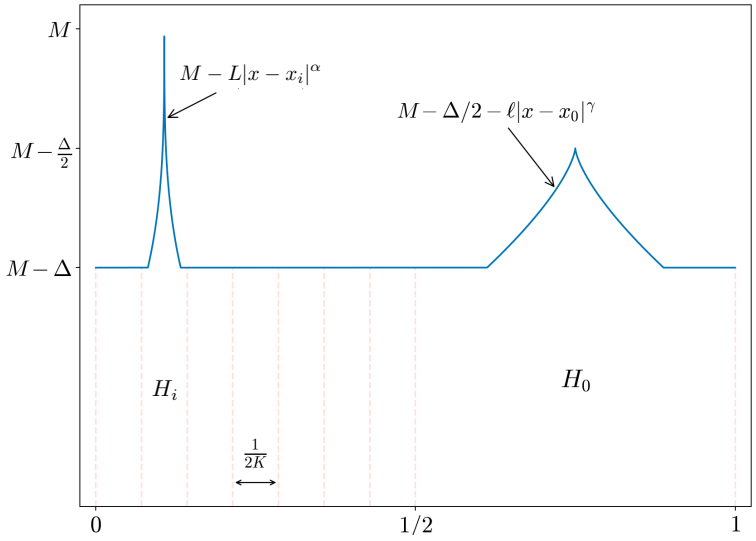


Figure 3.5.: Mean-payoff functions for the lower bound

Figure 3.5 illustrates how the ϕ_i's are defined : for $1 \leqslant i \leqslant K$, the function ϕ_i displays a peak of size Δ and of low regularity (L, α), localized in H_i, and another peak of size $\Delta/2$, of higher regularity (ℓ, γ) in H_0. The function ϕ_0 only has the peak of size $\Delta/2$ and regularity (ℓ, γ). We need to add requirements on the values of the parameters, to make sure the indeed functions belong to the appropriate regularity classes. These requirements are written out in the following lemma, which we prove later.

Lemma 3.4. *If* $(\Delta/L)^{1/\alpha} \leqslant 1/(4K)$ *then* $\phi_0 \in \mathcal{H}(\ell, \gamma)$*, and if* $(\Delta/(2\ell))^{1/\gamma} \leqslant 1/4$ *then* $\phi_i \in \mathcal{H}(L, \alpha)$ *for* $i \geqslant 1$.

Fix a given algorithm. The idea of the proof of the lower bound is to use the fact that if the player has low regret, that is, less than B, when the mean-payoff function is $\phi_0 \in \mathcal{H}(L, \alpha)$, then she has to play in H_0 often. This in turn constrains the amount of exploration she can afford, and limits her ability to find the maximum when the mean-payoff functions is ϕ_i for $i > 0$.

Canonical bandit model In this paragraph, we build the necessary setting for a rigorous development. The continous action space gives rise to measurability issues, and one should be particularly careful when handling changes of measure as we do here. Following Lattimore and Szepesvári [2020, Chap. 4.7, 14 (Ex.11) and 15 (Ex.8)], we build the canonical bandit model in order to apply the chain rule for Kullback-Leibler divergences rigorously. To our knowledge, this is seldom done carefully, the two notable exceptions being the above reference and Garivier et al. [2019]. We also use the notion of probability kernels in this paragraph; see Kallenberg [2006, Chap. 1 and 5] for a definition and properties.

Define a sequence of measurable spaces $\Omega_t = \prod_{s=1}^{t} \mathcal{X} \times \mathbb{R}$, together with their Borel σ-algebra (with the usual topology on $\mathcal{X} = [0, 1]$ and on $\mathbb{R}$). We call $h_t = (x_1, y_1, \ldots, x_t, y_t) \in \Omega_t$ a history up to time t. By an abuse of notation, we consider that $\Omega_t \subset \Omega_{t'}$ when $t \leqslant t'$.

An algorithm is a sequence $(K_t)_{1 \leqslant t \leqslant T}$ of (regular) probability kernels, with K_t from Ω_{t-1} to $\mathcal{X}$, modelling the choice of the arm at time t. By an abuse of notation, the first kernel K_1 is an arbitrary measure on $\mathcal{X}$, the law of the first arm picked. Define for each i another probability kernel modelling the reward obtained: $L_{i,t}$ from $\Omega_t \times \mathcal{X}$ to $\mathbb{R}$. We write it explicitly as :

$$L_{i,t}\big((x_1, y_1, \ldots, x_t), B\big) = \sqrt{\frac{2}{\pi}} \int_B e^{-2\left(x - \phi_i(x_t)\right)^2} \, \mathrm{d}x$$

These kernels define probability laws $\mathbb{P}_{i,t} = L_{i,t}(K_t \mathbb{P}_{i,t-1})$ over Ω_t. Doing so, we ensured that under $\mathbb{P}_{i,t}$ the coordinate random variables $X_t : \Omega_t \to \mathcal{X}$ and $Y_t : \Omega_t \to \mathbb{R}$), defined as $X_t(x_1, \dots, x_t, y_t) = x_t$ and $Y_t(x_1, \dots, x_t, y_t) = y_t$ are such that given X_t, the reward Y_t is distributed according to $\mathcal{N}\big(\phi_i(X_t), 1/4\big)$. Denote by $\mathbb{E}_i$ the expectation taken according to $\mathbb{P}_{i,t}$. We also index recall the pseudo-regret: $\overline{R}_{T,i} = TM(\phi_i) - \mathbb{E}_i\Big[\sum_{t=1}^T \phi_i(X_t)\Big]$.

A rewriting of the chain rule for Kullback-Leibler divergence with our notation would be (see Lattimore and Szepesvári [2020, Exercise 11 Chap. 14] for a proof)

Proposition (Chain rule). *Let Ω and Ω' be measurable subsets of $\mathbb{R}^d$ equipped with their natural σ-algebra. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions defined over Ω, and K and L be regular probability kernels from Ω to Ω' then*

$$\mathrm{KL}\left(K\mathbb{P}, L\mathbb{Q}\right) = \mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \int_\Omega \mathrm{KL}\big(K(\omega, \cdot), L(\omega, \cdot)\big)\, d\mathbb{P}(\omega)$$

The key assumptions are that Ω and Ω' are subspaces of $\mathbb{R}^d$, and that K and L satisfy measurability conditions, as they are regular kernels; these assumptions justify the heavy setting we introduced.

Under this setting, we may call to the chain rule twice to see that for any t:

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_0^t, \mathbb{P}_i^t\right) = \mathrm{KL}\left(L_{0,t}(K_t \mathbb{P}_0^{t-1}), L_{i,t}(K_t \mathbb{P}_i^{t-1})\right) \\
&= \mathrm{KL}\left(K_t \mathbb{P}_0^{t-1}, K_t \mathbb{P}_i^{t-1}\right) + \int_{\Omega_{t-1}\times\mathcal{X}} \mathrm{KL}\left(L_{0,t}(h_{t-1}, x_t, \cdot), L_{i,t}(h_{t-1}, x_t, \cdot)\right) \mathrm{d}K_t \mathbb{P}_0^{t-1}(h_{t-1}, x_t) \\
&= \mathrm{KL}\left(\mathbb{P}_0^{t-1}, \mathbb{P}_i^{t-1}\right) + \int_{\Omega_{t-1}\times\mathcal{X}} \mathrm{KL}\left(L_{0,t}(h_{t-1}, x_t, \cdot), L_{i,t}(h_{t-1}, x_t, \cdot)\right) \mathrm{d}K_t \mathbb{P}_0^{t-1}(h_{t-1}, x_t) \\
&= \mathrm{KL}\left(\mathbb{P}_0^{t-1}, \mathbb{P}_i^{t-1}\right) + \int_{\mathcal{X}} \mathrm{KL}\big(\mathcal{N}(\phi_0(x_t), 1/4), \mathcal{N}(\phi_i(x_t), 1/4)\big)\, \mathrm{d}\mathbb{P}_0^{t-1}(x_t) \\
&= \mathrm{KL}\left(\mathbb{P}_0^{t-1}, \mathbb{P}_i^{t-1}\right) + \mathbb{E}_0\big[\mathrm{KL}\big(\mathcal{N}(\phi_0(X_t), 1/4), \mathcal{N}(\phi_i(X_t), 1/4)\big)\big]
\end{aligned}$$

where the penultimate equality comes from the fact that the density of the kernel $L_{i,t-1}$ depends only on the last coordinate x_t, and is exactly that of a gaussian variable.

We obtain the KL decomposition by iterating T times,

$$\mathrm{KL}\left(\mathbb{P}_0^T, \mathbb{P}_i^T\right) = \mathbb{E}_0\left[\sum_{t=1}^T \mathrm{KL}\big(\mathcal{N}(\phi_0(X_t), 1/4), \mathcal{N}(\phi_i(X_t), 1/4)\big)\right]$$

Continuation of the proof Let us also define $N_{H_i}(T) = \sum_{t=1}^T \mathbb{1}_{\{X_t \in H_i\}}$ the number of times the algorithm selects an arm in H_i. The hypotheses ϕ_i were defined for the three following inequalities to hold. For all $i \geqslant 1$:

$$\overline{R}_{T,i} \geqslant \frac{\Delta}{2}\left(T - \mathbb{E}_i\big[N_{H_i}(T)\big]\right) = \frac{T\Delta}{2}\left(1 - \frac{\mathbb{E}_i\big[N_{H_i}(T)\big]}{T}\right), \tag{3.28}$$

$$\overline{R}_{T,0} \geqslant \frac{\Delta}{2}\sum_{i=1}^K \mathbb{E}_0\big[N_{H_i}(T)\big], \tag{3.29}$$

and

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_i^T) &= \mathbb{E}_0\left[\sum_{t=1}^T \mathrm{KL}\big(\mathcal{N}(\phi_0(X_t), 1/4), \mathcal{N}(\phi_i(X_t), 1/4)\big)\right] \\
&= \mathbb{E}_0\left[\sum_{t=1}^T 2\big(\phi_0(X_t) - \phi_i(X_t)\big)^2\right] \leqslant 2\,\mathbb{E}_0\big[N_{H_i}(T)\big]\Delta^2\,.
\end{aligned} \tag{3.30}$$

The first two inequalities come from the fact that, under $\mathbb{P}_i$, the player incurs an instantaneous regret of less than $\Delta/2$ whenever she picks an arm outside the optimal cell H_i. For the third inequality, first apply the chain rule to compute the Kullback-Leibler divergence, then the inequality is a consequence of the fact that ϕ_i and ϕ_0 differ only in H_i, and their difference is less than Δ.

We may now proceed with the calculations. By Lemma 3.3 applied to the random variable $N_{H_i}(T)/T$:

$$\frac{\mathbb{E}_i\big[N_{H_i}(T)\big]}{T} \leqslant \frac{\mathbb{E}_0\big[N_{H_i}(T)\big]}{T} + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_i^T)}{2}}\,. \tag{3.31}$$

We will now show that

$$\frac{1}{K}\sum_{i=1}^{K} \overline{R}_{T,i} \geqslant \frac{T\Delta}{2}\left(1 - \frac{1}{K} - \sqrt{\frac{\Delta\,\overline{R}_{T,0}}{K}}\right)\,. \tag{3.32}$$

Indeed by (in order) averaging (3.28) over $i = 1, \dots, K$, using (3.31), the concavity of $\sqrt{\cdot}$ and (3.30)

$$\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K} \overline{R}_{T,i} &\geqslant \frac{T\Delta}{2}\left(1 - \frac{1}{K}\sum_{i=1}^{K} \frac{\mathbb{E}_i\big[N_{H_i}(T)\big]}{T}\right) \\
&\geqslant \frac{T\Delta}{2}\left(1 - \frac{1}{K}\sum_{i=1}^{K} \frac{\mathbb{E}_0\big[N_{H_i}(T)\big]}{T} - \frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_i^T)}{2}}\right) \\
&\geqslant \frac{T\Delta}{2}\left(1 - \frac{1}{K} - \sqrt{\frac{1}{2K}\sum_{i=1}^{K}\mathrm{KL}(\mathbb{P}_0^T, \mathbb{P}_i^T)}\right) \\
&\geqslant \frac{T\Delta}{2}\left(1 - \frac{1}{K} - \sqrt{\frac{\Delta^2}{K}\sum_{i=1}^{K}\mathbb{E}_0\big[N_{H_i}(T)\big]}\right)\,.
\end{aligned}$$

This yields the claimed inequality (3.32) thanks to (3.29).

Let us assume for now that $K \geqslant 2$ and $\phi_0 \in \mathcal{H}(\ell, \gamma)$. Then by the assumption on the algorithm, $\overline{R}_{T,0} \leqslant B$, and therefore

$$\frac{1}{K}\sum_{i=1}^{K} \overline{R}_{T,i} \geqslant \frac{T\Delta}{2}\left(\frac{1}{2} - \sqrt{\frac{\Delta B}{K}}\right)\,. \tag{3.33}$$

To optimize this bound, we take Δ as large as possible, while still ensuring that $\sqrt{\Delta B/K}$ is small enough, e.g., less than $1/4$. Furthermore, we impose that the ϕ_i's belong to $\mathcal{H}(L, \alpha)$, i.e., by Lemma 3.4, that $(\Delta/L)^{1/\alpha} \leqslant 1/(4K)$. This leads to the choice

$$\Delta = c\,L^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)} \quad \text{and} \quad K = \left\lfloor \frac{1}{4}\left(\frac{\Delta}{L}\right)^{-1/\alpha} \right\rfloor = \left\lfloor \frac{c^{-1/\alpha}}{4}(LB)^{1/(\alpha+1)} \right\rfloor\,,$$

with $c = 1/128$.

Conclusion, assuming that $K \geqslant 2$ and $\phi_0 \in \mathcal{H}(\ell, \gamma)$ With this choice of parameters, we have by definition of Δ,

$$\Delta B = c\,(LB)^{1/(\alpha+1)}\,,$$

and by definition of K, since $K \geqslant (c^{-1/\alpha}/8)(LB)^{1/(\alpha+1)}$,

$$\frac{\Delta B}{K} \leqslant 8c^{1+1/\alpha}$$

hence, using $c^{1/(2\alpha)} \leqslant 1$

$$\sqrt{\frac{\Delta B}{K}} \leqslant 2\sqrt{2}c^{1/2+1/(2\alpha)} \leqslant 2\sqrt{2} \cdot 2^{-7/2} = \frac{1}{4}\,.$$

With this in hand, we may now go back to inequality (3.33) to see that

$$\frac{1}{K}\sum_{i=1}^{K} \overline{R}_{T,i} \geqslant \frac{T\Delta}{2}\left(\frac{1}{2} - \frac{1}{4}\right) \geqslant \frac{T\Delta}{8} = \frac{c}{8}\, TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}\,.$$

By the defintion of K, it is always true that $(\Delta/L)^{1/\alpha} \leqslant 1/(4K)$, and therefore, by Lemma 3.4, all the ϕ_i's automatically belong to $\mathcal{H}(L,\alpha)$. Therefore, for all i, we have $\sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T \geqslant \overline{R}_{T,i}$. Hence, recalling that $c = 1/128$,

$$\sup_{f\in\mathcal{H}(L,\alpha)} \overline{R}_T \geqslant \frac{1}{K}\sum_{i=1}^{K} \overline{R}_{T,i} \geqslant 2^{-10}\, TL^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}\,.$$

Regularity conditions on the mean-payoff functions ϕ_i We now check that $K \geqslant 2$, and that $\phi_0 \in \mathcal{H}(\ell,\gamma)$. Let us first focus on ϕ_0. By Lemma 3.4, it is enough to impose that $(\Delta/(2\ell))^{1/\gamma} \leqslant 1/4$, i.e., that

$$c\, L^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)}/(2\ell) \leqslant (1/4)^{\gamma}$$

that is,

$$L^{1/(\alpha+1)}B^{-\alpha/(\alpha+1)} \leqslant 2\ell(1/4)^{\gamma}/c = \ell\, 2^{1-2\gamma}c^{-1}\,,$$

i.e., when

$$LB^{-\alpha} \leqslant \ell^{1+\alpha}\, 2^{(1-2\gamma)(1+\alpha)}c^{-(1+\alpha)}$$

hence, replacing c by its value $c = 2^{-7}$, the next condition is sufficient to ensure the regularity of the hypothesis:

$$L \leqslant \ell^{1+\alpha}\, B^{\alpha}\, c^{-(1+\alpha)}\, 2^{(1+\alpha)(1-2\gamma)} = \ell^{1+\alpha}\, B^{\alpha}\, 2^{(1+\alpha)(8-2\gamma)}\,,$$

which is one of the two conditions in the statement of the theorem. For the bound to be valid, we must also make sure that $K \geqslant 2$:

$$\left\lfloor \left(\frac{c^{-1/\alpha}}{4}(LB)^{1/(\alpha+1)}\right)\right\rfloor \geqslant 2\,.$$

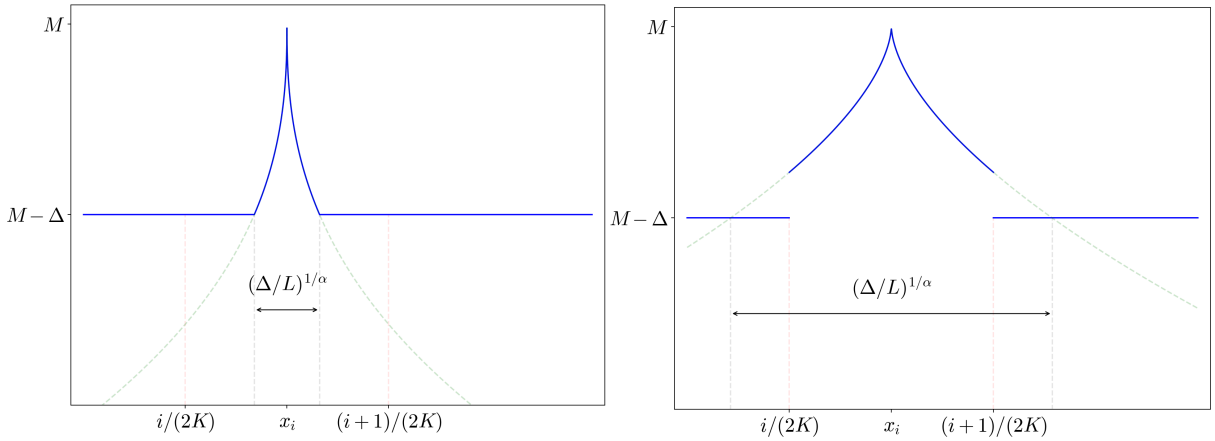
This condition is weaker than

$$\frac{c^{-1/\alpha}}{4}(LB)^{1/(\alpha+1)} \geqslant 3$$

which is equivalent to

$$L \geqslant c^{(\alpha+1)/\alpha}\, 12^{\alpha+1}B^{-1} = 2^{-7} \cdot 12 \cdot 2^{-6/\alpha}12^{\alpha}B^{-1}\,.$$

To ensure this, we require the stronger (but more readable) condition that $L \geqslant 2^{-3}12^{\alpha}B^{-1}$. □



(a) $(\Delta/L)^{1/\alpha} \leqslant 1/(4K)$ hence $\phi_i \in \mathcal{H}(L,\alpha)$ (b) $(\Delta/L)^{1/\alpha} > 1/(4K)$ hence $\phi_i \notin \mathcal{H}(L,\alpha)$

Figure 3.6.: ϕ_i is in $\mathcal{H}(L,\alpha)$ if it is everywhere above the green dotted curve $x \mapsto M - L\,|x - x_i|^\alpha$, that is, if the cell H_i has enough room to contain the whole peak of size Δ

Proof of Lemma 3.4. A good look at Figure 3.6 should convince the reader of the statement. We wish to make sure that the functions ϕ_i's satisfy (3.4), a Hölder condition around their maximum (and only around this maximum). Given the definition of the functions ϕ_i, we simply have to check that there is no discontinuity at the boundary of the cell H_i. We write out the details for $i > 0$ to remove any doubt; the same analysis can be carried to check that $\phi_0 \in \mathcal{H}(\ell,\gamma)$.

For $i > 0$, the function ϕ_i reaches its maximum at $x_i = (i - 1/2)/2K$, and the value of the maximum is M. Then for $x \in H_i$, by definition of ϕ_i:

$$\phi_i(x) = \max\left(M - \Delta, M - L\,|x_i - x|^\alpha\right) \geqslant M - L\,|x - x_i|^\alpha$$

thus

$$\phi_i(x_i) - \phi_i(x) = M - \phi_i(x) \leqslant L\,|x_i - x|^\alpha\ ,$$

Now consider $x \notin H_i$. Assume, as in the statement of the lemma, that $1/(4K) \geqslant (\Delta/L)^{1/\alpha}$. If x is outside of H_i, then since H_i is of half-width $1/4K$,

$$|x_i - x| \geqslant \frac{1}{4K} \geqslant \left(\frac{\Delta}{L}\right)^{1/\alpha} \tag{3.34}$$

and, by definition of ϕ_i, for all x (even for $x \in H_0$), $\phi_i(x) \geqslant M - \Delta$. Therefore, by (3.34),

$$\phi_i(x_i) - \phi_i(x) \leqslant \Delta \leqslant L\,|x_i - x|^\alpha\ .$$

For all values of x, the Hölder condition is satisfied and $\phi_i \in \mathcal{H}(L,\alpha)$.

For ϕ_0, the same calculations show that there is no jump at the boundary of $[1/2,1]$, of half-width $1/4$, when the peak is of height $\Delta/2$ and regularity (ℓ,γ) if $\big((\Delta/2)/\ell\big)\big)^{1/\gamma} \leqslant 1/4$. □

Chapter 4.

Adaptation to the range in K–armed bandits

Abstract

We consider stochastic bandit problems with K arms, each associated with a bounded distribution supported on the range $[m, M]$. We do not assume that the range $[m, M]$ is known and show that there is a cost for learning this range. Indeed, a new trade-off between distribution-dependent and distribution-free regret bounds arises, which, for instance, prevents from simultaneously achieving the typical $\ln T$ and $\sqrt{T}$ bounds. For instance, a $\sqrt{T}$ distribution-free regret bound may only be achieved if the distribution-dependent regret bounds are at least of order $\sqrt{T}$. We exhibit a strategy achieving the rates for regret indicated by the new trade-off.

This work was led in collaboration with Gilles Stoltz. The preprint Hadiji and Stoltz [2020] is currently under review.

Contents

4.1. Introduction

Virtually all articles on stochastic K–armed bandits either assume that distributions of the arms belong to some parametric family (often, one-dimensional exponential families) or to the non-parametric family of distributions supported on a known range $[m, M]$. Notable exceptions are discussed below.

We consider the second, non-parametric, framework (see Section 4.2) and show that the knowledge of the range $[m, M]$ is a crucial information. We do so by studying what may be achieved and what cannot be achieved anymore when this range is unknown and the strategies need to learn it. We call this problem scale-free regret minimization. In Section 4.3, we recall the standard distribution-dependent lower bound of Burnetas and Katehakis [1996], deriving its consequences on our problem. Our main result (in Section 4.4) is a trade-off between the scale-free distribution-dependent and distribution-free regret bounds that may be achieved; it is, for instance, impossible to simultaneously achieve scale-free distribution-dependent regret bounds of order $\ln T$ and scale-free distribution-free regret bounds of order $\sqrt{T}$, as simple strategies like UCB strategies (by Auer et al. [2002a]) do in the case of a known range. Our general trade-off indicates, for instance, that if one wants to keep the same $\sqrt{T}$ order of magnitude for the scale-free distribution-free regret bounds, then the best scale-free distribution-dependent rate that may be achieved is $\sqrt{T}$. We also provide (in Section 4.5) a strategy, based on exponential weights, that obtains optimal scale-free distribution-dependent and distribution-free regret bounds as indicated by the trade-off. We conclude the main body of the chapter with some numerical experiments illustrating the performance of our algorithms in scale-free regret minimization (in Section 4.6).

4.1.1. Literature review.

Closely related work. Optimal scale-free regret minimization under full monitoring is offered by the AdaHedge strategy by De Rooij et al. [2014], which we will use as a building block in in Section 4.5. The main difficulty in adaptation to the range is the adaptation to the upper end M (see Section 4.7); this is why Honda and Takemura [2015] could provide optimal $\ln T$ distribution-dependent regret bounds for payoffs lying in ranges of the form $(-\infty, M]$, with a known M. Lattimore [2017] considers models of distributions with a known bound on their kurtosis (a scale-free measure of the skewness of the distributions) and provides a scale-free algorithm based on the median-of-means estimators, with $\ln T$ distribution-dependent regret bounds. However, bounded bandits can have an arbitrarily high kurtosis, so our settings are not directly comparable (and we think that bounded distributions with an unknown range is a more natural assumption). Cowan and Katehakis [2015] study adaptation to the range but in the restricted case of uniform distributions; see also similar results by Cowan et al. [2018] for Gaussian distributions with unknown means and variances.

Adaptation to the effective range in adversarial bandits. Gerchinovitz and Lattimore [2016] show that it is impossible to adapt to the so-called effective range in adversarial bandits. A

sequence of rewards has effective range smaller than b if for all rounds t, rewards $y_{t,a}$ at this round all lie in an interval of the form $[m_t, M_t]$ with $M_t - m_t \leqslant b$. The lower bound they exhibit relies on a sequence of changing intervals of fixed size. This problem is thus different from our setting. See also positive results (upper bounds) by Cesa-Bianchi and Shamir [2018] for adaptation to the effective range.

Adaptation to the variance. Audibert et al. [2009] consider a variant of UCB called UCB-V, which adapts to the unknown variance. Its analysis assumes that rewards lie in a known range $[0, M]$. The results crucially use Bernstein's inequality (see, for instance, Reminder 4.3 in Appendix 4.A.2 for a statement of the latter); as Bernstein's inequality holds for random variables with supports in $[-\infty, M]$, the analysis of UCB-V might perhaps be extended to this case as well. Deviation bounds in Bernstein's inequality contain two terms, a main term scaling with the standard deviation, and a remainder term, scaling with M. This remainder term, which seems harmless, is a true issue when M is not known, as indicated by the results of the present chapter.

Other criteria. Wei and Luo [2018], Zimmert and Seldin [2019], Bubeck et al. [2018], and many more, provide strategies for adversarial bandits with rewards in a known range, say $[0, 1]$, and adapting to additional regularity in the data, like small variations or stochasticity of the data.

4.2. Settings: stochastic bandits and bandits for oblivious individual sequences

We describe the bandit settings considered: stochastic bandits, the setting of main interest, and bandits for oblivious individual (adversarial) sequences, a setting leading to stronger regret upper bounds.

4.2.1. Stochastic bandits with bounded and possibly signed rewards

$K \geqslant 2$ arms are available. We denote by $[K]$ the set $\{1, \dots, K\}$ of arms. With each of the arm a is associated a probability distribution ν_a lying in some known model $\mathcal{D}$; a model is a set of probability distributions over $\mathbb{R}$ with a first moment. The models of interest in this chapter are discussed below. A bandit problem over $\mathcal{D}$ is a K–vector of probability distributions in $\mathcal{D}$: we denote it by $\underline{\nu} = (\nu_a)_{a\in[K]}$. The player knows $\mathcal{D}$ but not $\underline{\nu}$. As is standard in this setting, we denote by $\mu_a = \mathrm{E}(\nu_a)$ the mean payoff provided by an arm a. An optimal arm and the optimal mean payoff are respectively given by $a^\star \in \mathrm{argmax}_{a\in[K]}\, \mu_a$ and $\mu^\star = \max_{a\in[K]} \mu_a$. Finally, $\Delta_a = \mu^\star - \mu_a$ denotes the gap of an arm a.

The online learning game goes as follows: at round $t \geqslant 1$, the player picks an arm $A_t \in [K]$, possibly at random according to a probability distribution $p_t = (p_{t,a})_{a\in[K]}$ based on an auxiliary randomization U_{t-1}, and then receives and observes a reward Z_t drawn independently at random according to the distribution ν_{A_t}, given A_t. More formally, a strategy of the player is a sequence of mappings from the observations to the action set, $(U_0,\, Z_1, U_1,\, \dots,\, Z_{t-1}, U_{t-1}) \mapsto A_t$, where $U_0, U_1, \dots$ are i.i.d. random variables independent from all other random variables and distributed according to a uniform distribution over $[0, 1]$. At each given time $T \geqslant 1$, we measure the performance of a strategy through its expected regret:

$$R_T(\underline{\nu}) = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Z_t\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right] = \sum_{a=1}^{K} \Delta_a\, \mathbb{E}\big[N_a(T)\big]\,, \tag{4.1}$$

where we used the tower rule for the first equality and defined $N_a(T)$ as the number of times arm a was pulled between time rounds 1 and T.

Doob's optional skipping (see Doob [1953, Chapter III, Theorem 5.2, p. 145] for the original reference, see also Chow and Teicher [1988, Section 5.3] for a more recent reference) indicates that we may assume that i.i.d. sequences of rewards $(Y_{t,a})_{t\geqslant 1}$ are drawn beforehand, independently at random, for each arm a and that the obtained payoff at round $t \geqslant 1$ given the choice A_t equals $Z_t = Y_{t,A_t}$. We will use this second formulation in the rest of the chapter as it is the closest to the one of oblivious individual sequences described in Section 4.2.2.

Models: bounded rewards with unknown range. For a given range $[m, M]$, where $m < M$ are two real numbers (not necessarily nonnegative), we denote by $\mathcal{D}_{m,M}$ the set of probability distributions supported on $[m, M]$. Then, the model corresponding to distributions with a bounded but unknown range is the union of all such $\mathcal{D}_{m,M}$:

$$\mathcal{D}_{-,+} = \bigcup_{\substack{m,M\in\mathbb{R}\\ m<M}} \mathcal{D}_{m,M}\,. \tag{4.2}$$

4.2.2. Oblivious individual sequences (oblivious adversarial bandits)

In the setting of (fully) oblivious individual sequences (see Cesa-Bianchi and Lugosi [2006], Audibert and Bubeck [2009]), a range $[m, M]$ is set by the environment, where m, M are real numbers (not necessarily nonnegative), and the environment picks beforehand a sequence $y_1, y_2, \ldots$ of reward vectors in $[m, M]^K$. We denote by $y_t = (y_{t,a})_{a\in[K]}$ the components of these vectors. The online learning game starts only then: at each round $t \geqslant 1$, the player picks an arm $A_t \in [K]$, possibly at random according to a probability distribution $p_t = (p_{t,a})_{a\in[K]}$ based on an auxiliary randomization U_{t-1}, and then receives and observes y_{t,A_t}. More formally, a strategy of the player is a sequence of mappings from the observations to the action set, $(U_0,\, y_{1,A_1}, U_1,\, \ldots,\, y_{t-1,A_{t-1}}, U_{t-1}) \mapsto A_t$, where $U_0, U_1, \ldots$ are i.i.d. random variables independent from all other random variables and distributed according to a uniform distribution over $[0, 1]$. At each given time $T \geqslant 1$, denoting $y_{1:T} = (y_1, \ldots, y_T)$, we measure the performance of a strategy through its expected regret:

$$R_T(y_{1:T}) = \max_{a\in[K]} \sum_{t=1}^{T} y_{t,a} - \mathbb{E}\left[\sum_{t=1}^{T} y_{t,A_t}\right], \tag{4.3}$$

where all randomness lies in the choice of the arms A_t only (as rewards are fixed beforehand).

Conversion of upper/lower bounds from one setting to the other. Note that (by the tower rule for the right-most equality) for all $m < M$ and for all $\underline{\nu}$ in $\mathcal{D}_{m,M}$,

$$\begin{aligned} R_T(\underline{\nu}) = \max_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,a}\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,A_t}\right] &\leqslant \mathbb{E}\left[\max_{a\in[K]} \sum_{t=1}^{T} Y_{t,a} - \sum_{t=1}^{T} Y_{t,A_t}\right] = \mathbb{E}\big[R_T(Y_{1:T})\big] \\ &\leqslant \sup_{y_{1:T}\ \mathrm{in}[m,M]^K} R_T(y_{1:T})\,. \end{aligned}$$

In particular, lower bounds on the regret for stochastic bandits are also lower bounds on the regret for oblivious adversarial bandits, and strategies designed for oblivious adversarial bandits obtain the same regret bounds for stochastic bandits when the individual payoffs y_{t,A_t} in their definition are replaced with the stochastic payoffs Y_{t,A_t}.

4.2.3. Scale-free regret bounds: rates for adaptation to the unknown range

Regret scales with the range length $M - m$, thus regret bounds involve a multiplicative factor $M - m$. We therefore consider such bounds divided by the scale factor $M - m$ and call them scale-free regret bounds. We denote by $\mathbb{N}$ the set of natural integers; (rates on) regret bounds will be given by functions $\Phi : \mathbb{N} \to [0, +\infty)$.

Definition 4.1 (Distribution-free bounds). *A strategy for stochastic bandits, respectively, for oblivious individual sequences, is adaptive to the unknown range of payoffs with a scale-free distribution-free regret bound $\Phi : \mathbb{N} \to [0, +\infty)$ if for all real numbers $m < M$, the strategy ensures, without the knowledge of m and M:*

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{m,M}, \quad \forall T \geqslant 1, \qquad R_T(\underline{\nu}) \leqslant (M - m)\, \Phi(T)\,,$$

respectively, $$\forall y_1, y_2, \dots \text{ in } [m, M]^K, \quad \forall T \geqslant 1, \qquad R_T(y_{1:T}) \leqslant (M - m)\, \Phi(T)\,.$$

The notion of distribution-dependent regret bounds for adaptation to the range can obviously only be defined for stochastic bandits. It does not add much to the classical notion of distribution-dependent rates on regret bounds, as the scale factor $M - m$ does not appear in the definition; it merely ensures that the strategy is not informed of the range.

Definition 4.2 (Distribution-dependent bounds). *A strategy for stochastic bandits is adaptive to the unknown range of payoffs with a distribution-dependent rate $\Phi : \mathbb{N} \to [0, +\infty)$ if for all real numbers $m < M$, the strategy ensures, without the knowledge of m and M:*

$$\forall \underline{\nu} \text{ in } \mathcal{D}_{m,M}, \qquad \limsup_{T \to +\infty} \frac{R_T(\underline{\nu})}{\Phi(T)} < +\infty\,.$$

Put differently, the strategy ensures that $\limsup R_T(\underline{\nu})/\Phi(T) < +\infty$ for all $\underline{\nu} \in \mathcal{D}_{-,+}$.

4.3. Distribution-dependent lower bounds for adaptation to the range

Any scale-free distribution-free regret bound $\Phi_{\text{free}}(T)$ is larger than the optimal distribution-free regret bound on a known range. Auer et al. [2002b] provided a lower bound $(1/20) \min\bigl\{\sqrt{KT}, T\bigr\}$ on the regret of any strategy against individual sequences in $[0, 1]^K$, thus for bandit problems in $\mathcal{D}_{0,1}$. Therefore, we also have $\Phi_{\text{free}}(T) \geqslant (1/20) \min\bigl\{\sqrt{KT}, T\bigr\}$. We show in Section 4.5 a scale-free distribution-free regret bound of order $\sqrt{KT \ln K}$, which thus matches the lower bound up to a $\sqrt{\ln K}$ factor.

The situation is different for distribution-dependent bounds, where the typical $\ln T$ order of magnitude cannot be achieved when the range is unknown: all uniformly fast convergent strategies on $\mathcal{D}_{-,+}$ (see Definition 4.3 below) are such that, for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{-,+}$ with at least one suboptimal arm,

$$\liminf_{T \to +\infty} \frac{R_T(\underline{\nu})}{\ln T} = +\infty\,. \tag{4.4}$$

However, any rate $\varphi(T) \gg \ln T$ may be achieved thanks to a simple upper-confidence bound [UCB] strategy.

Before we expand on these two statements, we remind the reader of the "classical" results, for an abstract model $\mathcal{D}$ and then, for the model $\mathcal{D}_{m,M}$ corresponding to payoff distributions with a known range $[m, M]$.

Definition 4.3. *A strategy is uniformly fast convergent on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, it achieves a subpolynomial regret bound, that is, $R_T(\underline{\nu})/T^\alpha \to 0$ for all $(\alpha, 1]$.*

A lower bound on the distribution-dependent rates that such a strategy may achieve is provided by a general result of Lai and Robbins [1985] and Burnetas and Katehakis [1996] (see also its rederivation by Garivier et al. [2019]). It involves a quantity defined as an infimum of Kullback-Leibler divergences: we recall that for two probability distributions ν, ν' defined on the same probability space $(\Omega, \mathcal{F})$,

$$\mathrm{KL}(\nu, \nu') = \begin{cases} \displaystyle\int_\Omega \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)\mathrm{d}\nu & \text{if } \nu \ll \nu', \\ +\infty & \text{otherwise,} \end{cases}$$

where $\nu \ll \nu'$ means that ν is absolutely continuous with respect to ν' and $\mathrm{d}\nu/\mathrm{d}\nu'$ then denotes the Radon-Nikodym derivative. Now, for any probability distribution ν, any real number x, and any model $\mathcal{D}$, we define

$$\mathcal{K}_{\inf}(\nu, x, \mathcal{D}) = \inf\bigl\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mathrm{E}(\nu') > x\bigr\}\,,$$

where by convention, the infimum of an empty set equals $+\infty$ and where we denoted by $\mathrm{E}(\nu')$ the expectation of ν'. The quantity $\mathcal{K}_{\inf}(\nu, x, \mathcal{D})$ can be null. With the usual measure-theoretic conventions, in particular, $0/0 = 0$, we then have the following lower bound.

Reminder 4.1. *For all models $\mathcal{D}$, for all uniformly fast convergent strategies on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$,*

$$\liminf_{T\to+\infty} \frac{R_T(\underline{\nu})}{\ln T} \geqslant \sum_{a\in[K]} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\,.$$

When the range $[m, M]$ of payoffs is known, i.e., when the model is $\mathcal{D}_{m,M}$, there exist strategies achieving the lower bound of Reminder 4.1, like the DMED strategy of Honda and Takemura [2011, 2015] or the KL–UCB strategy of Cappé et al. [2013] and Garivier et al. [2018]. (This can even be extended to the case of semi-bounded only rewards with a known upper bound on the payoffs, as is discussed in details in Appendix 4.B.2.)

No logarithmic regret distribution-dependent regret bound under adaptation to the range. Now, the lower bound in Reminder 4.1 cannot be achieved any more when the range is not known, that is, when we consider the model $\mathcal{D}_{-,+}$ of bounded distributions with unknown range. Actually, the proof reveals that the important fact is that the upper end of the payoff range is unknown. The impossibility result also holds for models $\mathcal{D}_{m,+}$ of bounded distributions with unknown upper end on the range and known lower end m on the range, for some fixed $m \in \mathbb{R}$:

$$\mathcal{D}_{m,+} = \bigcup_{\substack{M\in\mathbb{R},\\ m<M}} \mathcal{D}_{m,M}\,. \tag{4.5}$$

Theorem 4.1. *All uniformly fast convergent strategies on $\mathcal{D}_{-,+}$ are such that, for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{-,+}$ with at least one suboptimal arm,*

$$\liminf_{T\to+\infty} \frac{R_T(\underline{\nu})}{\ln T} = +\infty\,.$$

The same result holds for all models $\mathcal{D}_{m,+}$, where $m \in \mathbb{R}$.

Strategies that are adaptive to the range thus cannot get rates Φ for distribution-dependent regret bounds on the regret of the order of $\ln T$ in Definition 4.2. A similar phenomenon was

discussed by Lattimore [2017] in the case of stochastic bandits with sub-Gaussian distributions. It turns out that any rate Φ such that $\Phi(T) \gg \ln T$ may be achieved, through a simple upper-confidence bound [UCB] strategy, as also discussed by Lattimore [2017]; see further details after the proof.

Proof. We fix $m \in \mathbb{R}$ and provide the proof for $\mathcal{D}_{m,+}$. Given Reminder 4.1 and since we assumed that at least one arm a is suboptimal, i.e., is associated with a gap $\Delta_a = \mu^\star - \mu_a > 0$, it is necessary and sufficient to show that $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) = 0$, where $\nu_a \in \mathcal{D}_{m,+}$.

We have in particular $\mu_a \geqslant m$. We use the same construction as in the proof of Theorem 4.2. Let $\nu'_\varepsilon = (1-\varepsilon)\nu_a + \varepsilon\delta_{\mu_a + 2\Delta_a/\varepsilon}$ for $\varepsilon \in (0,1)$: it is a bounded probability distribution, with lower end of support larger than m, that is, $\nu'_\varepsilon \in \mathcal{D}_{m,+}$. For ε small enough, $\mu_a + 2\Delta_a/\varepsilon$ lies outside of the bounded support of ν_a. In that case, the density of ν_a with respect to ν'_ε is given by $1/(1-\varepsilon)$ on the support of ν_a (and 0 elsewhere), so that

$$\mathrm{KL}\big(\nu_a, \nu'_\varepsilon\big) = \ln\bigg(\frac{1}{1-\varepsilon}\bigg).$$

Moreover, $\mathrm{E}\big(\nu'_\varepsilon\big) = (1-\varepsilon)\mu_a + \varepsilon\big(\mu_a + 2\Delta_a/\varepsilon\big) = \mu_a + 2\Delta_a = \mu^\star + \Delta_a > \mu^\star$. Therefore, by definition of $\mathcal{K}_{\inf}$ as an infimum,

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) \leqslant \mathrm{KL}\big(\nu_a, \nu'_\varepsilon\big) = \ln\bigg(\frac{1}{1-\varepsilon}\bigg).$$

This upper bound holds for all $\varepsilon > 0$ small enough and thus shows that we actually have $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,+}) = 0$.

The exact same construction and proof can be performed in the case of $\mathcal{D}_{-,+}$, without the need of indicating that the lower end of the support of ν'_ε is larger than m. □

UCB with an increased exploration rate adapts to the range The lower bound of Theorem 4.1 does not prevent distribution-dependent rates for adaptation that are arbitrarily larger than a logarithm. Consider UCB with indexes of the form

$$\widehat{\mu}_a(t) + \sqrt{\frac{\varphi(t)}{N_a(t)}} \qquad \text{where} \qquad \frac{\varphi(t)}{\ln t} \to +\infty$$

and where $\widehat{\mu}_a(t)$ denotes the empirical average of payoffs obtained till round t when playing arm a. Following the analysis of Lattimore [2017] in the case of Gaussian bandits with unknown variances, it can be shown that such a UCB is adaptive to the unknown range of payoffs with a distribution-dependent rate φ. However the trick used here is purely asymptotic and gives up on finite-time guarantees.

4.4. Regret lower bounds for adaptation to the range

We now show that under an adaptivity assumption that is stronger than uniform fast convergence and takes finite-time guarantees into account, the distribution-dependent regret becomes polynomial in T.

4.4.1. Simultaneous scale-free distribution-free and distribution-dependent lower bounds

When the range $[m, M]$ of the payoffs is known, it is possible to simultaneously achieve optimal distribution-free bounds (of order $\sqrt{KT}$) and optimal distribution-dependent bounds (of order

$\ln T$ with the optimal constant given by infima of Kullback-Leibler divergences); see the KL-UCB-switch strategy by Garivier et al. [2018]. Put differently, when the range of payoffs is known, one can achieve optimal (asymptotic) distribution-dependent regret bounds while not sacrificing finite-time guarantees. Simpler strategies like UCB strategies (see Auer et al. [2002a]) also simultaneously achieve regret bounds of similar $\ln T$ and $\sqrt{T}$ orders of magnitude but with suboptimal constants and/or dependencies on K.

This is not possible anymore when the range of payoffs is unknown.

To show this, we consider in this section algorithms enjoying distribution-free scale-free regret bounds and show that they suffer up to a $\Omega(\sqrt{T})$ distribution-dependent rate for adaptation to the range. Actually, the theorem below shows that there is a trade-off between the finite-time guarantees (the distribution-free scale-free regret bounds) and the asymptotic problem-dependent rates (the distribution-dependent rates for adaptation) that can be achieved. We recall that these concepts were defined in Section 4.2.3. The proof actually provides a finite-time (but messy) lower bound on $R_T(\underline{\nu})\big/\big(T/\Phi_{\mathrm{free}(T)}\big)$.

Theorem 4.2. *Any strategy with a Φ_{free} distribution-free scale-free regret bound satisfying $\Phi_{free} \ll T$ may only achieve distribution-dependent rates Φ_{dep} for adaptation satisfying $\Phi_{dep}(T) \geqslant T/\Phi_{free}(T)$. More precisely, the regret of such a strategy is lower bounded as: for all $\underline{\nu}$ in $\mathcal{D}_{-,+}$,*

$$\liminf_{T\to\infty} \frac{R_T(\underline{\nu})}{T/\Phi_{free(T)}} \geqslant \frac{1}{4}\sum_{a=1}^{K} \Delta_a \,.$$

The optimal distribution-free scale-free regret bounds $\Phi_{\mathrm{free}}(T)$ are of order $\sqrt{T}$ (as follows from the lower bound indicated at the beginning of Section 4.4 and from the upper bound of Section 4.5). The distribution-dependent rates $\Phi_{\mathrm{dep}}(T)$ of strategies achieving this optimal distribution-free scale-free rate are therefore larger than $\sqrt{T}$. More generally, there is a trade-off between the two rates: to force faster distribution-dependent rates for adaptation, one must suffer worsened distribution-free scale-free rates for adaptation. (The latter range between the optimal $\sqrt{KT}$ rate and the trivial T rate.)

4.4.2. Proof of Theorem 4.2

We follow a standard proof technique introduced by Lai and Robbins [1985] and Burnetas and Katehakis [1996] and recently revisited by Garivier et al. [2019]. We fix some bandit problem $\underline{\nu}$ in $\mathcal{D}_{-,+}$ and construct an alternative bandit problem $\underline{\nu}'$ in $\mathcal{D}_{-,+}$ by modifying the distribution of a single suboptimal arm a to make it optimal (which is always possible, as there is no bound on the upper end on the ranges of the payoffs in the model). We apply a fundamental inequality that links the expectations of the numbers of times $N_a(T)$ that a is pulled under $\underline{\nu}$ and $\underline{\nu}'$. We then substitute inequalities stemming from the definition of distribution-free scale-free regret bounds Φ_{dep}, and the result follows by rearranging all inequalities.

Step 1: Alternative bandit problem. The lower bound is trivial (it equals 0) when all arms of $\underline{\nu}$ are optimal. We therefore assume that at least one arm is suboptimal and fix such an arm a. For some $\varepsilon \in [0,1]$ to be defined later by the analysis, we introduce the alternative problem $\underline{\nu}' = (\nu'_k)_{k\in[K]}$ with $\nu'_k = \nu_k$ for $j \neq a$ and $\nu'_a = (1-\varepsilon)\nu_a + \varepsilon\delta_{\mu_a+2\Delta_a/\varepsilon}$. This distribution ν'_a has a bounded range, so that $\underline{\nu}'$ lies indeed in $\mathcal{D}_{-,+}$. The expectation of ν'_a equals $\mu'_a = \mu_a + 2\Delta_a = \mu^\star + \Delta_a > \mu^\star$. Thus, a is the only optimal arm in $\underline{\nu}'$. Finally, for ε small enough, $\mu_a + 2\Delta_a/\varepsilon$ lies outside of the bounded support of ν_a. In that case, the density of ν_a with respect to ν'_ε is given by $1/(1-\varepsilon)$ on the support of ν_a (and 0 elsewhere), so that $\mathrm{KL}(\nu_a, \nu'_a) = \ln\big(1/(1-\varepsilon)\big)$.

Step 2: Application of a fundamental inequality. We denote by $\mathrm{kl}(p,q)$ the Kullback-Leibler divergence between Bernoulli distributions with parameters p and q. We also index expectations

in the rest of the proof by the bandit problem they are relative to: for instance, $\mathbb{E}_{\underline{\nu}}$ denotes the expectation of a random variable when the ambiant randomness is given by the bandit problem $\underline{\nu}$. The fundamental inequality for lower bounds on the regret of stochastic bandits (Garivier et al. [2019], Section 2, Equation 6), which is based on the chain rule for Kullback-Leibler divergence and on a data-processing inequality for expectations of $[0,1]$–valued random variables, reads:

$$\mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T}, \frac{\mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]}{T}\right) \leqslant \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\ \mathrm{KL}(\nu_a, \nu'_a) = \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\ln\big(1/(1-\varepsilon)\big)\,.$$

Now, since $u \in (-\infty, 1) \mapsto -u^{-1}\ln(1-u)$ is increasing, we have $\ln\big(1/(1-\varepsilon)\big) \leqslant \varepsilon(\ln 2)/2$ for $\varepsilon \leqslant 1/2$. For all $(p,q) \in [0,1]^2$ and with the usual measure-theoretic conventions,

$$\mathrm{kl}(p,q) = \underbrace{p\ln p + q\ln q}_{\geqslant -\ln 2} + \underbrace{p\ln\frac{1}{q}}_{\geqslant 0} + (1-p)\ln\frac{1}{1-q} \geqslant (1-p)\ln\frac{1}{1-q} - \ln 2\,,$$

so that, putting all inequalities together, we have proved

$$\left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T}\right)\ln\left(\frac{1}{1 - \mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]/T}\right) - \ln 2 \leqslant \frac{\ln 2}{2}\varepsilon\,\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,. \tag{4.6}$$

So far, we only imposed the constraint $\varepsilon \in [0, 1/2]$.

Step 3: Inequalities stemming from the definition of distribution-free scale-free regret bounds. We denote by $[m, M]$ a range containing the supports of all distributions of $\underline{\nu}$. By definition of Φ_{free}, given that a is a suboptimal arm (i.e., $\Delta_a > 0$):

$$\Delta_a\,\mathbb{E}_{\underline{\nu}}[N_a(T)] \leqslant R_T(\underline{\nu}) \leqslant (M-m)\,\Phi_{\mathrm{free}}(T)\,.$$

Because of ν'_a, the distributions of $\underline{\nu}'$ have supports within the range $[m, M_\varepsilon]$, where we denoted $M_\varepsilon = \max\{M,\, \mu_a + 2\Delta_a/\varepsilon\}$. For $\underline{\nu}'$, by definition of Φ_{free}, and given that all gaps Δ'_k are larger than the gap $\Delta'_a = \mu'_a - \mu^\star = \Delta_a$ between the unique optimal a and the second best arms (which were the optimal arms of $\underline{\nu}$),

$$\Delta_a\big(T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]\big) = \Delta'_a\big(T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]\big) \leqslant \sum_{j\neq a}\Delta'_j\,\mathbb{E}_{\underline{\nu}'}[N_k(T)] = R_T(\underline{\nu}') \leqslant (M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)\,.$$

By rearranging the two inequalities above, we get

$$1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T} \geqslant 1 - \frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a} \qquad \text{and} \qquad 1 - \frac{\mathbb{E}_{\underline{\nu}'}\big[N_a(T)\big]}{T} \leqslant \frac{(M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}\,,$$

thus, after substitution into (4.6),

$$\left(1 - \frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}\right)\ln\left(\frac{T\Delta_a}{(M_\varepsilon - m)\,\Phi_{\mathrm{free}}(T)}\right) - \ln 2 \leqslant \frac{\ln 2}{2}\varepsilon\,\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,. \tag{4.7}$$

Step 4: Final calculations. We take $\varepsilon = \varepsilon_T = \alpha^{-1}\,\Phi_{\mathrm{free}}(T)/T$ for some constant $\alpha > 0$; we will pick $\alpha = 1/8$. By the assumption $\Phi_{\mathrm{free}}(T) \ll T$, we have $\varepsilon_T \leqslant 1/2$ as needed for T large enough, as well as $M_{\varepsilon_T} = \mu_a + 2\Delta_a/\varepsilon_T = \mu_a + 2\alpha\Delta_a T/\Phi_{\mathrm{free}}(T)$. Substituting these values into (4.7), a finite-time lower bound on the quantity of interest is finally given by

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{T/\Phi_{\mathrm{free}}(T)} \geqslant \frac{2\alpha}{\ln 2}\left(-\ln 2 + \left(1 - \underbrace{\frac{(M-m)\,\Phi_{\mathrm{free}}(T)}{T\Delta_a}}_{\to 0}\right)\ln\left(\underbrace{\frac{T\Delta_a}{2\alpha\Delta_a T + (\mu_a - m)\Phi_{\mathrm{free}}(T)}}_{\to 1/(2\alpha)}\right)\right).$$

It entails the asymptotic lower bound

$$\liminf_{T\to+\infty} \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{T/\Phi_{\mathrm{free}}(T)} \geqslant \frac{2\alpha}{\ln 2}\big(\ln(1/\alpha) - 2\ln 2\big) = \frac{1}{4}$$

for the choice $\alpha = 1/8$. The claimed result follows by adding these lower bounds for each suboptimal arm a, with a factor Δ_a, following the rewriting (4.1) of the regret.

4.5. Quasi-optimal regret bounds for range adaptation based on AdaHedge

When the range of payoffs is known, Auer et al. [2002b] use exponential weights (Hedge) on estimated payoffs and with extra-exploration (mixing with the uniform distribution) to achieve a regret bound of order $\sqrt{KT\ln K}$. Actually, it is folklore knowledge that the extra-exploration is unnecessary when regret bounds are considered only in expectation, as is the case in the present chapter.

When the range of payoffs is unknown, we consider a self-tuned version called AdaHedge (De Rooij et al. [2014], see also earlier work by Cesa-Bianchi et al. [2007]) and do add extra-exploration. The latter is not detrimental, given the trade-off between the distribution-free and distribution-dependent bounds discussed in the previous section; we actually achieve that trade-off. Algorithm 4.1 is stated in the case of adversarial oblivious learning, but to use it with stochastic payoffs, if suffices to replace y_{t,A_t} with Y_{t,A_t}. It relies on a payoff estimation scheme, which we discuss now.

In Algorithm 4.1, some initial exploration lasting K rounds is used to get a rough idea of the location of the payoffs and to center the estimates used at an appropriate location. Following by Auer et al. [2002b]), we consider, for all rounds $t \geqslant K+1$ and arms $a \in [K]$,

$$\widehat{y}_{t,a} = \frac{y_{t,A_t} - C}{p_{t,a}} \mathbb{1}_{\{A_t=a\}} + C \qquad \text{where} \qquad C \stackrel{\text{def}}{=} \frac{1}{K}\sum_{s=1}^{K} y_{s,s}\,. \tag{4.8}$$

Note that all $p_{t,a} > 0$ for Algorithm 4.1 due to the use of exponential weights. As proved by Auer et al. [2002b], these estimates are (conditionally) unbiased. Indeed, the distributions q_t and p_t (as well as the constant C) are measurable functions of the information $H_{t-1} = (U_0, y_{1,A_1}, U_1, \ldots, y_{t-1,A_{t-1}})$ available at the beginning of round $t \geqslant K+1$, and the arm A_t is drawn independently at random according to p_t based on an auxiliary randomization denoted by U_{t-1}. Therefore, given that the payoffs are oblivious, the conditional expectation of $\widehat{y}_{t,a}$ with respect to H_{t-1} amounts to integrating over the randomness given by the random draw $A_t \sim p_t$: for $t \geqslant K+1$,

$$\mathbb{E}\big[\widehat{y}_{t,a} \,\big|\, H_{t-1}\big] = \frac{y_{t,a} - C}{p_{t,a}}\,\mathbb{P}\big(A_t = a \,\big|\, H_{t-1}\big) + C = \frac{y_{t,a} - C}{p_{t,a}}\,p_{t,a} + C = y_{t,a}\,. \tag{4.9}$$

These estimators are bounded: assuming that all $y_{t,a}$, thus also C, belong to the range $[m, M]$, and given that the distributions p_t were obtained by a mixing with the uniform distribution, with weight γ_t, we have $p_{t,a} \geqslant \gamma_t/K$, and therefore,

$$\forall t \geqslant K+1, \quad \forall a \in [K], \qquad \big|\widehat{y}_{t,a} - C\big| \leqslant \frac{|y_{t,a} - C|}{p_{t,a}} \leqslant \frac{M-m}{\gamma_t/K}\,. \tag{4.10}$$

Remark. *Algorithm 4.1 is invariant by affine changes (translations and/or multiplications by positive factors) of the payoffs, as AdaHedge (see De Rooij et al. [2014, Theorem 16]) and the payoff estimation scheme* (4.8) *so are. This is key for adaptation to the range.*

Algorithm 4.1 AdaHedge for K–armed bandits, with extra-exploration

Input: a sequence $(\gamma_t)_{t\geqslant 1}$ in $[0,1]$ of extra-exploration rates; a payoff estimation scheme
for rounds $t = 1, \dots, K$ **do**
  Draw arm $A_t = t$
  Get and observe the payoff $y_{t,t}$
end for
AdaHedge initialization: $\eta_{K+1} = +\infty$ and $q_{K+1} = (1/K, \dots, 1/K) \overset{\text{def}}{=} \mathbf{1}/K$
for rounds $t = K+1, \dots$ **do**
  Define p_t by mixing q_t with the uniform distribution according to $p_t = (1-\gamma_t)q_t + \gamma_t \mathbf{1}/K$
  Draw an arm $A_t \sim p_t$ (independently at random according to the distribution p_t)
  Get and observe the payoff y_{t,A_t}
  Compute estimates $\widehat{y}_{t,a}$ of all payoffs with the payoff estimation scheme considered
  Compute the mixability gap $\delta_t \geqslant 0$ based on the distribution q_t and on these estimates:

$$\delta_t = -\sum_{a=1}^{K} q_{t,a}\,\widehat{y}_{t,a} + \frac{1}{\eta_t}\ln\left(\sum_{a=1}^{K} q_{t,a} \mathrm{e}^{\eta_t \widehat{y}_{t,a}}\right), \qquad \text{with} \quad \underbrace{\delta_t = \sum_{a=1}^{K} q_{t,a}\,\widehat{y}_{t,a} + \max_{a\in[K]} \widehat{y}_{t,a}}_{\text{when } \eta_t = +\infty}$$

  Compute the learning rate $\eta_{t+1} = \left(\sum_{s=K+1}^{t} \delta_s\right)^{-1} \ln K$
  Define q_{t+1} component-wise as

$$q_{t+1,a} = \exp\left(\eta_{t+1} \sum_{s=K+1}^{t} \widehat{y}_{a,s}\right) \Bigg/ \sum_{k=1}^{K} \exp\left(\eta_{t+1} \sum_{s=K+1}^{t} \widehat{y}_{k,s}\right)$$

end for

4.5.1. Distribution-free scale-free regret analysis

Theorem 4.3. *AdaHedge for K–armed bandits (Algorithm 4.1) with a non-increasing extra-exploration (γ_t) smaller than $1/2$ and the estimation scheme given by (4.8) ensures that for all bounded ranges $[m, M]$, for all oblivious individual sequences $y_1, y_2, \dots$ in $[m, M]^K$, for all $T \geqslant 1$,*

$$R_T(y_{1:T}) \leqslant 3(M-m)\sqrt{KT\ln K} + 5(M-m)\frac{K\ln K}{\gamma_T} + (M-m)\sum_{t=K+1}^{T} \gamma_t\,.$$

Proof sketch. We provide only a sketch of proof and refer the reader to 4.A for a complete, detailed and commented proof. A direct application of the AdaHedge regret bound (Lemma 3 and Theorem 6 of De Rooij et al. [2014]), bounding the variance terms of the form $\mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$ by $\mathbb{E}\big[(X-C)^2\big]$, ensures that

$$\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{\substack{t\geqslant K+1 \\ a\in[K]}}^{T} q_{t,a}\,\widehat{y}_{t,a} \leqslant 2\sqrt{\sum_{\substack{t\geqslant K+1 \\ a\in[K]}} q_{t,a}\big(\widehat{y}_{t,a} - C\big)^2 \ln K} + \frac{M-m}{\gamma_T/K}\left(2 + \frac{4}{3}\ln K\right).$$

We take expectations, use the definition of the p_t in terms of the q_t in the left-hand side, and

apply Jensen's inequality in the right-hand side to get

$$\mathbb{E}\left[\max_{k\in[K]}\sum_{t=K+1}^{T}\widehat{y}_{t,k}-\sum_{t=K+1}^{T}\overbrace{\sum_{a=1}^{K}p_{t,a}\,\widehat{y}_{t,a}}^{=y_{t,A_t}}+\sum_{t=K+1}^{T}\gamma_t\overbrace{\sum_{a=1}^{K}(1/K-q_{t,a})\,\widehat{y}_{t,a}}^{\mathbb{E}[\ldots]\in[m-M,M-m]}\right]$$
$$\leqslant 2\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K}\mathbb{E}\Big[q_{t,a}\big(\widehat{y}_{t,a}-C\big)^2\Big]\ln K}+\frac{M-m}{\gamma_T/K}\left(2+\frac{4}{3}\ln K\right).$$

Since $p_{t,a}\geqslant(1-\gamma_t)q_{t,a}$ with $\gamma_t\leqslant 1/2$ by assumption on the extra-exploration rate, we have the bound $q_{t,a}\leqslant 2p_{t,a}$. Together with standard calculations similar to (4.9), we have

$$\mathbb{E}\Big[q_{t,a}\big(\widehat{y}_{t,a}-C\big)^2\Big]\leqslant 2\,\mathbb{E}\Big[p_{t,a}(\widehat{y}_{t,a}-C)^2\,\Big|\,H_{t-1}\Big]=2\,\mathbb{E}\left[\frac{(y_{t,A_t}-C)^2}{p_{t,a}}\mathbb{1}_{\{A_t=a\}}\right]=2\underbrace{(y_{t,a}-C)^2}_{\leqslant(M-m)^2}.$$

The proof is concluded by collecting all bounds and by taking care of the first K rounds. □

Straightforward calculations then lead to the following consequence of Theorem 4.3.

Corollary 4.1. *Fix a parameter $\alpha\in(0,1)$. AdaHedge for K–armed bandits (Algorithm 4.1) with the extra-exploration*

$$\gamma_t=\min\Big\{1/2,\ \sqrt{5(1-\alpha)K\ln K}/t^{\alpha}\Big\}$$

and the estimation scheme given by (4.8) *ensures that for all bounded ranges $[m,M]$, for all oblivious individual sequences $y_1,y_2,\ldots$ in $[m,M]^K$, for all $T\geqslant 1$,*

$$R_T(y_{1:T})\leqslant\left(3+\frac{5}{\sqrt{1-\alpha}}\right)(M-m)\sqrt{K\ln K}\;T^{\max\{\alpha,1-\alpha\}}+10(M-m)K\ln K\,.$$

In particular, for $\alpha=1/2$, the bound $7(M-m)\sqrt{TK\ln K}+10(M-m)K\ln K$ holds.

Proof. We have, first,

$$\sum_{t=K+1}^{T}\gamma_t\leqslant\sqrt{5(1-\alpha)K\ln K}\sum_{t=K+1}^{T}t^{-\alpha}\leqslant\sqrt{5(1-\alpha)K\ln K}\int_0^T\frac{1}{t^{\alpha}}\,\mathrm{d}t=\sqrt{\frac{5K\ln K}{1-\alpha}}T^{1-\alpha},$$

second, using the definition of γ_T as a minimum,

$$\frac{K\ln K}{\gamma_T}\leqslant\frac{K\ln K}{1/2}+\frac{T^{\alpha}K\ln K}{\sqrt{5(1-\alpha)K\ln K}}=2K\ln K+\sqrt{\frac{K\ln K}{5(1-\alpha)}}\,T^{\alpha},$$

and third, $\sqrt{T}\leqslant T^{\max\{\alpha,1-\alpha\}}$, so that the regret bound of Theorem 4.3 may be further bounded by

$$R_T(y_{1:T})\leqslant(M-m)\sqrt{K\ln K}\left(3+2\sqrt{\frac{5}{1-\alpha}}\right)T^{\max\{\alpha,1-\alpha\}}+10(M-m)K\ln K\,.$$

The claimed bound is obtained by bounding $2\sqrt{5}$ by 5. □

This value $\alpha=1/2$ is the best one to consider if one is only interested in a distribution-free bound (i.e., one is not interested in the distribution-dependent rates for the regret).

4.5.2. Distribution-dependent regret analysis, and discussion of the trade-off

For $\alpha \in [1/2, 1)$, Algorithm 4.1 tuned as in Corollary 4.1 is adaptive to the unknown range of payoffs with a distribution-free scale-free regret bound

$$\Phi^{\mathrm{AH}}_{\mathrm{free}}(T) = \left(3 + \frac{5}{\sqrt{1-\alpha}}\right)\sqrt{K \ln K}\; T^{\alpha} + 10K \ln K \tag{4.11}$$

for oblivious individual sequences thus also for stochastic bandits, with the same regret bound. (The superscript AH in $\Phi^{\mathrm{AH}}_{\mathrm{free}}$ stands for AdaHedge.) The trade-off stated in Theorem 4.2 indicates that the best possible distribution-dependent rate for adaptation to the unknown range is determined by $T/\Phi^{\mathrm{AH}}_{\mathrm{free}}(T)$, which is of order $T^{1-\alpha}$. It indicates, more precisely, that for all $\underline{\nu}$ in $\mathcal{D}_{-,+}$,

$$\liminf_{T\to\infty} \frac{R_T(\underline{\nu})}{T/\Phi^{\mathrm{AH}}_{\mathrm{free}(T)}} \geqslant \frac{1}{4}\sum_{a=1}^{K} \Delta_a \,.$$

The following theorem shows that this best possible distribution-dependent rate is indeed achieved and quantifies the gap between the distribution-dependent constants at hand: they differ by two multiplicative factors, a numerical factor of $4 \times 12/(1-\alpha)$ and a $\ln K$ factor.

Theorem 4.4. *Consider Algorithm 4.1 tuned as in Corollary 4.1, for $\alpha \in [1/2, 1)$. For all distributions $\nu_1, \ldots, \nu_K$ in $\mathcal{D}_{-,+}$,*

$$\limsup_{T\to\infty} \frac{R_T(\underline{\nu})}{T/\Phi^{AH}_{free(T)}} \leqslant \frac{12 \ln K}{1-\alpha} \sum_{a=1}^{K} \Delta_a \,. \tag{4.12}$$

The proof is provided in 4.A. It follows quite closely that of Theorem 3 in Seldin and Lugosi [2017], where the authors study a variant of the Exp3 algorithm of Auer et al. [2002b] for stochastic rewards. It consists, in our setting, in showing that the number of times the algorithm chooses suboptimal arms is almost only determined by the extra-exploration. Our proof is simpler as we aim for cruder bounds. The main technical difference and issue to solve lies in controlling the learning rates η_t, which heavily depend on data in our case.

4.6. Numerical experiments

We describe some numerical experiments on synthetic data to illustrate the performance of the new algorithms introduced compared to earlier approaches; we focus on how algorithms adapt to the scale of payoffs.

Five (families of) algorithms are considered. The first algorithm compared is vanilla UCB (with a $2\ln T$ exploration factor, as in the original reference by Auer et al. [2002a]) and only adapt it to take the range $[m, M]$ of payoffs into account, by adding a $M - m$ factor in front of the upper confidence bound (see details below). We also compare AdaHedge for bandits and another strategy, alluded at in Section 4.7 and to be described in details in Appendices 4.B.2 and 4.C.4, called AdaFTRL with 1/2–Tsallis entropy, a generalization of the INF strategy of Audibert and Bubeck [2009]. As the latter was introduced to handle losses (nonpositive payoffs), we will consider such nonpositive payoffs in our setting. It turns out that AdaHedge for bandits can be slightly improved in this case (see Appendices 4.B.2 and 4.C.3), by centering estimates at $C = 0$. Finally, we also add a simple follow-the-leader strategy (referred to as FTL; i.e., a strategy picking at each round the arm with best payoff estimate so far) and the random strategy (i.e., picking at each round an arm uniformly at random). FTL and the random strategies will exhibit undesirable performance similar to the ones of incorrectly tuned instances of UCB (respectively, with too small and too large a parameter σ).

Stochastic setting: bandit problems considered. We consider stochastic bandit problems $\underline{\nu}^{(\alpha)}$ indexed by a scale parameter $\alpha \in \{0.01, 0.1, 10, 100\}$. More precisely, $\underline{\nu}^{(\alpha)} = (\nu_a^{(\alpha)})_{a\in[K]}$ with $K = 30$ arms, each associated with a uniform distribution defined by

$$\nu_a^{(\alpha)} = \begin{cases} \mathrm{Unif}\big([\,-\alpha,\, 0]\big) & \text{if } a = 1, \\ \mathrm{Unif}\big([\,-1.2\,\alpha,\, -0.2\,\alpha]\big) & \text{if } a \neq 1, \end{cases}$$

so that all distributions are commonly supported on $[m, M] = [-1.2\,\alpha,\ 0]$, with arm 1 being the unique optimal arm. Given the scale values $M - m = 1.2\alpha$ obtained for the ranges $[m, M]$ as α varies, we consider four instances of UCB, with respective upper confidence bounds

$$\widehat{\mu}_a(t) + 1.2\,\sigma\sqrt{\frac{2\ln T}{N_a(t)}}\,, \qquad \text{for} \qquad \sigma \in \{0.01, 0.1, 10, 100\}\,,$$

where $N_a(t)$ is the number of times arm a was pulled up to round t and $\widehat{\mu}_a(t)$ denotes the empirical average of payoffs obtained for arm a when it was played.

Experimental setting. Each algorithm is run $N = 300$ times, on a time horizon $T = 100\,000$. We plot estimates of the rescaled regret $R_T(\underline{\nu}^{(\alpha)})/\alpha$ to have a meaningful comparison between the bandit problems.

These estimates are constructed as follows. We denote by

$$\mu_a^{(\alpha)} = \begin{cases} -\alpha/2 & \text{if } a = 1 \\ -0.7\,\alpha & \text{if } a \neq 1 \end{cases}$$

the mean of arm a in $\underline{\nu}^{(\alpha)}$. We index the arms picked in the n–th run by an additional subscript n, so that $A_{T,n}$ refers to the arm picked by some strategy at time t in the n–th run. The expected regret of a given strategy can be rewritten as

$$R_T(\underline{\nu}^{\alpha}) = T\max_{a\in[K]}\mu_a^{(\alpha)} - \mathbb{E}\left[\sum_{t=1}^{T}\mu_{A_t}^{(\alpha)}\right] = -T\alpha/2 - \mathbb{E}\left[\sum_{t=1}^{T}\mu_{A_t}^{(\alpha)}\right]$$

and is estimated by

$$\widehat{R}_T(\alpha) = \frac{1}{N}\sum_{n=1}^{N}\widehat{R}_T(\alpha, n) \qquad \text{where} \qquad \widehat{R}_T(\alpha, n) = -T\alpha/2 - \sum_{t=1}^{T}\mu_{A_{t,n}}^{(\alpha)}\,.$$

On Figure 4.1 we therefore plot the estimates $\widehat{R}_T(\alpha)/\alpha$ of the rescaled regret as solid lines. The shaded areas correspond to ± 2 standard errors of the sequences $\big(\widehat{R}_T(\alpha, n)/\alpha\big)_{n\in[N]}$.

Complexity. The time complexity of FTL and of the instances of UCB lies only in the update of the payoff estimate of the selected arm and in the choice of the next arm. AdaHedge for bandits has a higher runtime due to the additional cost of computing the distributions q_t and p_t over the arms and the mixability gaps. AdaFTRL for bandits with 1/2–Tsallis entropy is the most time consuming of our algorithms, as it requires twice solving an optimization problem: once for the computation of the distributions over the arms and once for the mixability gaps; see Section 4.C.4 for specific details on the said optimization problem and hints on an efficient solution thereof.

The memory complexity of all algorithms considered here is constant and scales linearly with K. The algorithms only need to keep in memory a vector of (cumulative or average) payoffs estimates and (for some) the cumulative mixability gaps.

Table 4.1.: Average runtimes of the (families of) algorithms considered, measured in seconds per run; as a reminder, we performed $N = 300$ runs for each algorithm.

Random play	FTL	UCB family	Bandit AdaHedge	Tsallis–AdaFTRL
$X = 1.51$ s /run	$1.7X$	$1.7X$	$7.9X$	$32.4X$

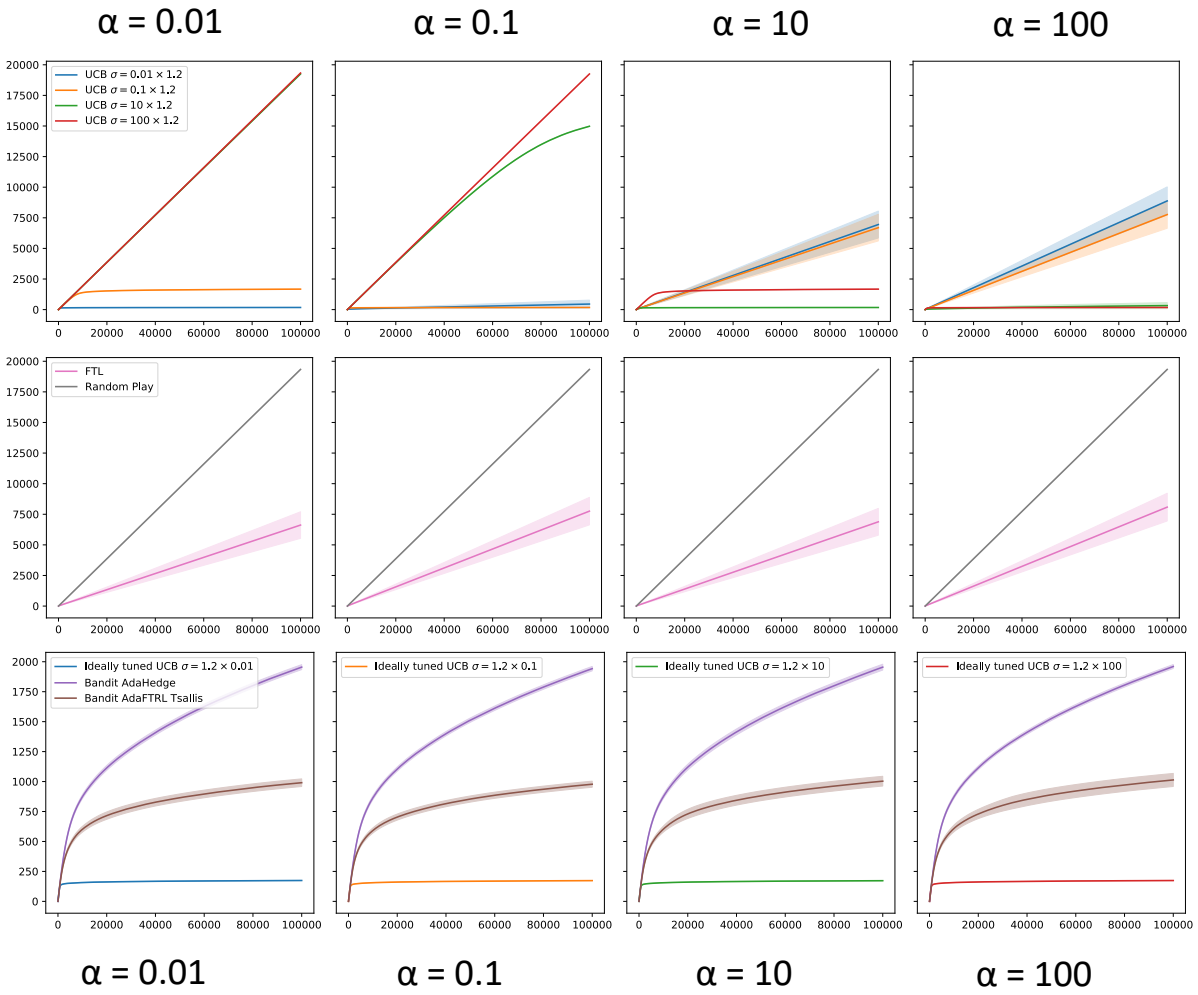


Figure 4.1.: Comparison of the rescaled regrets of various strategies over bandit problems $\underline{\nu}^{\alpha}$, where α ranges in $\{0.01, 0.1, 10, 100\}$. Each algorithm was run $N = 300$ times on every problem for $T = 100\,000$ time steps. Solid lines report the values of the estimated rescaled regrets, while shaded areas correspond to ± 2 standard errors of the estimates.

All experiments were designed in Python, using the NumPy and joblib libraries, and were run on a standard laptop computer (with an Intel Core i5 processor). The code and setup for these experiments were only moderately optimized for computational efficiency. We display the average runtimes of all algorithms in Table 4.1; they are provided only for illustration and could certainly be significantly improved .

Discussion of the results. A first observation is that, as expected, our algorithms (see the third lines of Figure 4.1) are unaffected by the scale of the problems (up to a minor numerical stability issue discussed below). They yield favorable results (note that the range of the y–axis for the third line is smaller than the ranges in the first two lines), with AdaFTRL with 1/2–Tsallis entropy exhibiting a better performance than AdaHedge for bandits (our theoretical bounds reflect this, see Appendices 4.B and 4.C).

UCB tuned with the correct scale obtains the best results overall, which is consistent with the folklore knowledge that UCB performs well in practice. However, and this is our major second observation, the performance of UCB worsens dramatically when the scale is misspecified. When UCB is run with a scale parameter σ that is too small, it behaves similarly to FTL, incurring linear regret with extreme variance. When the scale parameter σ is too large, UCB is essentially playing at random and incurs linear regret too.

We conclude this section by discussing a minor issue of numerical stability: the error bars of for the expected regret of AdaFTRL with 1/2–Tsallis entropy seem to increase slightly with the scale α (while in theory they are independent of α). This is probably due to larger numerical errors associated with the approximate solutions of the optimization problems discussed in Section 4.C.4.

A final note: UCB with estimated range. For the sake of completeness, we indicate that a version of UCB estimating the range, i.e., considering indices of the form

$$\widehat{\mu}_a(t) + \widehat{r}_t \sqrt{\frac{2\ln T}{N_a(t)}}\,,$$

where $\widehat{r}_t$ estimates the range $M - m$ as

$$\widehat{r}_t = \max_{s \leqslant t} Y_{A_s,s} - \min_{s \leqslant t} Y_{A_s,s}\,,$$

obtained an excellent performance on our simulations (the same as the optimally tuned version of UCB). We were however unable to provide theoretical guarantees that match our lower bounds. This is why we do not discuss this natural algorithm in the present chapter.

4.7. Extensions present in the appendix

One known end on the payoff range. It is folklore knowledge that there is a difference in nature between dealing with nonnegative payoffs (gains) or dealing with nonpositive payoffs (losses) for regret minimization under bandit monitoring; see Cesa-Bianchi and Lugosi [2006, Remark 6.5, page 164] for an early reference and Kwon and Perchet [2016] for a more complete literature review. Actually, 0 plays no special role, the issue is rather whether one end of the payoff range is known. What follows is detailed in 4.B.

Known lower end m on the payoff range. In that case we deal (up to a translation) with gains. This knowledge does not provide any advantage. Indeed, the impossibility results of Section 4.4 still hold, namely, no $\ln T$ rate may be achieved for scale-free distribution-dependent regret bounds, as in (4.4), and a trade-off exists between scale-free distribution-free and distribution-dependent regret bounds (Theorem 4.2 holds).

Known upper end M on the payoff range. In that case we deal (up to a translation) with losses, also known as semi-bounded rewards. The results of Section 4.4 do not hold anymore. The DMED strategy of Honda and Takemura [2015] achieves the optimal asymptotic distribution-dependent regret bound, of order $\ln T$. We also recover some classical results: the INF strategy of Audibert and Bubeck [2009] may be extended to provide a scale-free distribution-free regret bound of order $\sqrt{KT}$, and the AdaHedge strategy does not need any mixing with the uniform distribution to achieve the bound of Theorem 4.3.

Linear bandits. The techniques developed for adaptation to the range in Section 4.5 may be generalized to deal with (oblivious) adversarial linear bandits, see details in 4.D.

Appendix for Chapter 4

We provide the following additions and extensions to the core results described in the main body of the chapter.

Appendix 4.A provides the complete proofs of the results of Section 4.5, namely, the ones of Theorem 4.3, and Theorem 4.4.

Appendix 4.B studies whether the adaptation results described in the main body of the chapter in the case of an unknown payoff range $[m, M]$ still hold when only one end of this range is unknown. It turns out that when m is known but M is unknown, achieving a distribution-dependent bound for adaptation to the range of order $\ln T$ is still impossible, and the trade-off between scale-free distribution-free and distribution-dependent regret bounds still holds (Theorem 4.2 holds). The picture is completely different when M is known but m is unknown, and improved scale-free distribution-free regret bounds can be provided.

Appendix 4.C provides the statements and proofs of some technical results alluded in earlier appendices: the full-information regret bound for AdaHedge, needed in the complete proof of Theorem 4.3 in Appendix 4.A, as well as the improved scale-free distribution-free regret bounds in the case the upper end M of the payoff range is known. All these results rely on a self-tuned version of a follow-the-regularized-leader (FTRL) strategy called AdaFTRL.

Appendix 4.D deals with adaptation to the range for (oblivious) adversarial linear bandits.

4.A. Complete proofs of the results of Section 4.5

We provide here complete proofs for Theorem 4.3 and Theorem 4.4, in this order.

4.A.1. Proof of Theorem 4.3

In Algorithm 4.1, for time steps $t \geqslant K+1$, the weights q_t are obtained by using the AdaHedge algorithm of De Rooij et al. [2014] on the payoff estimates $\widehat{y}_{t,a}$. AdaHedge is designed for the case of a full monitoring (not a bandit monitoring), but the use of these estimates emulates a full monitoring. Section 2.2 of De Rooij et al. [2014] (see also an earlier analysis by Cesa-Bianchi et al. [2007]) ensures the bound stated next in Reminder 4.2. For the sake of completeness, we rederive this bound in Appendix 4.C.2. We call pre-regret the quantity at hand in Reminder 4.2: it corresponds to some regret defined in terms of the payoff estimates.

Reminder 4.2 (Application of Lemma 3 and Theorem 6 of De Rooij et al. [2014]). *For all sequences of payoff estimates $\widehat{y}_{t,a}$ lying in some bounded real-valued interval, denoted by $[b,B]$, for all $T \geqslant K+1$, the pre-regret of AdaHedge satisfies*

$$\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T} \sum_{a=1}^{K} q_{t,a}\, \widehat{y}_{t,a} \leqslant 2 \sum_{t=K+1}^{T} \delta_t$$

$$\text{where} \qquad \sum_{t=K+1}^{T} \delta_t \leqslant \underbrace{\sqrt{\sum_{t=K+1}^{T} \sum_{a=1}^{K} q_{t,a} \left(\widehat{y}_{t,a} - \sum_{k\in[K]} q_{t,k}\, \widehat{y}_{t,k} \right)^2 \ln K}}_{\leqslant \sqrt{\sum_{t=K+1}^{T} \sum_{a=1}^{K} q_{t,a} (\widehat{y}_{t,a}-c)^2 \ln K} \quad \text{for any } c\in\mathbb{R}} + (B-b) \left(1 + \frac{2}{3} \ln K \right)$$

and AdaHedge does not require the knowledge of $[b,B]$ to achieve this bound.

The bound of Reminder 4.2 will prove itself particularly handy for three reasons: first, it is valid for signed payoffs (payoffs in $\mathbb{R}$); second, it is adaptive to the range of payoffs; third, the right-hand side looks at first sight not intrinsic enough a bound (as it also depends on the weights q_t) but we will see later that this dependency is particularly useful.

We recall that we start the summation in Reminder 4.2 at $t = K+1$ because the AdaHedge algorithm is only started at this time, after the initial exploration. The bound holding "for any $c \in \mathbb{R}$" is obtained by a classical bound on the variance.

Proof of Theorem 4.3. We deal with the contribution of the initial exploration by using the inequality $\max(u+v) \leqslant \max u + \max v$, together with the fact that $y_{t,a} - y_{t,A_T} \leqslant M - m$ for any $a \in [K]$:

$$R_T(y_{1:T}) \leqslant \underbrace{\max_{a\in[K]} \sum_{t=1}^{K} y_{t,a} - \mathbb{E}\left[\sum_{t=1}^{K} y_{t,A_t} \right]}_{\leqslant K(M-m)} + \max_{a\in[K]} \sum_{t=K+1}^{T} y_{t,a} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t} \right]. \tag{4.13}$$

We now transform the pre-regret bound of Reminder 4.2, which is stated with the distributions q_t, into a pre-regret bound with the distributions p_t; we do so while substituting the bounds $B = C + KM/\gamma_T$ and $b = C + Km/\gamma_T$ implied by (4.10) and the fact that (γ_t) is non-increasing,

and by using the definition $q_{t,a} = p_{t,a} - \gamma_t(1/K - q_{t,a})$ for all $a \in [K]$:

$$\begin{aligned}&\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} + \sum_{t=K+1}^{T} \gamma_t \sum_{a=1}^{K} (1/K - q_{t,a})\,\widehat{y}_{t,a} \leqslant 2\sum_{t=K+1}^{T} \delta_t \\ &\text{where} \qquad \sum_{t=K+1}^{T} \delta_t \leqslant \sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K} + \frac{(M-m)K}{\gamma_T}\left(1 + \frac{2}{3}\ln K\right).\end{aligned} \tag{4.14}$$

As noted by Auer et al. [2002b], by the very definition (4.8) of the estimates,

$$\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} = y_{t,A_t}\,.$$

By (4.9), the tower rule and the fact that q_t is H_{t-1}–measurable, on the one hand, and the fact that the expectation of a maximum is larger than the maximum of expectations, on the other hand, the left-hand side of the first inequality in (4.14) thus satisfies

$$\begin{aligned}&\mathbb{E}\left[\max_{k\in[K]} \sum_{t=K+1}^{T} \widehat{y}_{t,k} - \sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}\,\widehat{y}_{t,a} + \sum_{t=K+1}^{T} \gamma_t \sum_{a=1}^{K} (1/K - q_{t,a})\,\widehat{y}_{t,a}\right] \\ &\geqslant \max_{k\in[K]} \sum_{t=K+1}^{T} y_{t,k} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t}\right] + \sum_{t=K+1}^{T} \gamma_t \Bigg(\underbrace{\sum_{a=1}^{K} y_{t,a}/K}_{\in[m,M]} - \underbrace{\sum_{a=1}^{K} \mathbb{E}\big[q_{t,a}\big]\, y_{t,a}}_{\in[m,M]}\Bigg) \\ &\geqslant \max_{k\in[K]} \sum_{t=K+1}^{T} y_{t,k} - \mathbb{E}\left[\sum_{t=K+1}^{T} y_{t,A_t}\right] - (M-m)\sum_{t=1}^{T}\gamma_t\,.\end{aligned}$$

As for the right-hand side of the second inequality in (4.14), we first note that by definition (see line 4 in Algorithm 4.1), $p_{t,a} \geqslant (1-\gamma_t)q_{t,a}$ with $\gamma_t \leqslant 1/2$ by assumption on the extra-exploration rate, so that $q_{t,a} \leqslant 2p_{t,a}$; therefore, by substituting first this inequality and then by using Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} q_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K}\right] &\leqslant \sqrt{2}\;\mathbb{E}\left[\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} p_{t,a}(\widehat{y}_{t,a} - C)^2 \ln K}\right] \\ &\leqslant \sqrt{2}\sqrt{\sum_{t=K+1}^{T}\sum_{a=1}^{K} \mathbb{E}\Big[p_{t,a}(\widehat{y}_{t,a} - C)^2\Big] \ln K}\,.\end{aligned} \tag{4.15}$$

Standard calculations (see Auer et al. [2002b] again) show, similarly to (4.9), that for all $a \in [K]$,

$$\mathbb{E}\Big[p_{t,a}(\widehat{y}_{t,a} - C)^2 \,\Big|\, H_{t-1}\Big] = \mathbb{E}\left[\frac{(y_{t,A_t} - C)^2}{p_{t,a}}\mathbb{1}_{\{A_t = a\}}\right] = (y_{t,a} - C)^2 \leqslant (M-m)^2\,,$$

where the last inequality comes from (4.10). By the tower rule, the same upper bound holds for the (unconditional) expectation. Therefore, taking the expectation of both sides of (4.14) and collecting all bounds together, we proved so far

$$R_T(y_{1:T}) \leqslant \underbrace{2\sqrt{2}}_{\leqslant 3}(M-m)\sqrt{KT\ln K} + (M-m)\frac{K\ln K}{\gamma_T}\underbrace{\left(\frac{2+\gamma_T}{\ln K} + \frac{4}{3}\right)}_{\leqslant 5} + (M-m)\sum_{t=K+1}^{T}\gamma_t\,,$$

where we used $\gamma_T \leqslant 1/2$ and $\ln K \geqslant \ln 2$ as $K \geqslant 2$. □

4.A.2. Proof of Theorem 4.4

Proof of Theorem 4.4. Given the decomposition (4.1) of the regret, it is necessary and sufficient to upper bound the expected number of times $\mathbb{E}[N_a(t)]$ any suboptimal arm a is drawn, where by definition of Algorithm 4.1,

$$\mathbb{E}[N_a(t)] = 1 + \mathbb{E}\left[\sum_{t=K+1}^{T}\left((1-\gamma_t)q_{t,a} + \frac{\gamma_t}{K}\right)\right] \leqslant 1 + \sum_{t=K+1}^{T}\mathbb{E}[q_{t,a}] + \frac{1}{K}\sum_{t=K+1}^{T}\gamma_t\,.$$

We show below (and this is the main part of the proof) that

$$\sum_{t=K+1}^{T}\mathbb{E}[q_{t,a}] = \mathcal{O}(\ln T)\,. \tag{4.16}$$

The proof of Corollary 4.1] shows in particular that

$$\frac{1}{K}\sum_{t=K+1}^{T}\gamma_t \leqslant \sqrt{\frac{5\ln K}{(1-\alpha)K}}\,T^{1-\alpha}\,.$$

Substituting the value (4.11) of $\Phi^{\mathrm{AH}}_{\mathrm{free}(T)}$ and using the decomposition (4.1) of $R_T(\underline{\nu})$ into $\sum \Delta_a\,\mathbb{E}[N_a(t)]$ then yield

$$\frac{R_T(\underline{\nu})}{T/\Phi^{\mathrm{AH}}_{\mathrm{free}(T)}} \leqslant \sum_{a\in[K]}\Delta_a\sqrt{\frac{5\ln K}{(1-\alpha)K}}\left(3 + \frac{5}{\sqrt{1-\alpha}}\right)\sqrt{K\ln K}\big(1+o(1)\big) + \mathcal{O}\left(\frac{\ln T}{T^{1-\alpha}}\right),$$

from which the stated bound follows, via the crude inequality $3\sqrt{5}\sqrt{1-\alpha}+5 \leqslant 12$.

Structure of the proof of (4.16). Let $a^\star$ denote an optimal arm. By definition of $q_{t,a}$ and by lower bounding a sum of exponential terms by any of the summands, we get

$$q_{t,a} = \frac{\exp\left(\eta_t\sum_{s=K+1}^{t-1}\widehat{y}_{t,a}\right)}{\sum_{k=1}^{K}\exp\left(\eta_t\sum_{s=K+1}^{t-1}\widehat{y}_{t,k}\right)} \leqslant \exp\left(\eta_t\sum_{t=K+1}^{t-1}(\widehat{y}_{t,a} - \widehat{y}_{t,a^\star})\right).$$

Then, by separating cases, depending on whether $\sum_{t=K+1}^{t-1}(\widehat{y}_{t,a} - \widehat{y}_{t,a^\star})$ is smaller or larger than $-(t-1-K)\Delta_a/2$, and by remembering that the probability $q_{t,a}$ is always smaller than 1, we get

$$\begin{aligned}\sum_{t=K+1}^{T}\mathbb{E}[q_{t,a}] \leqslant & \sum_{t=K+1}^{T}\mathbb{E}\left[\exp\left(-\eta_t\frac{(t-1-K)\Delta_a}{2}\right)\right] \\ & + \sum_{t=K+1}^{T}\mathbb{P}\left[\sum_{s=K+1}^{t-1}(\widehat{y}_{s,a} - \widehat{y}_{s,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2}\right].\end{aligned} \tag{4.17}$$

We show that the sums in the right-hand side of (4.17) are respectively $\mathcal{O}(1)$ and $\mathcal{O}(\ln T)$.

First sum in the right-hand side of (4.17). Given the definition of the learning rates (see the statement of Algorithm 4.1), namely,

$$\eta_t = \ln K \Big/ \sum_{s=K+1}^{t-1}\delta_s\,, \tag{4.18}$$

we are interested in upper bounds on the sum of the δ_s. Such upper bounds were already derived in the proof of Theorem 4.3; the second inequality in (4.14) together with the bound $q_{t,a} \leqslant 2p_{t,a}$ stated in the middle of the proof immediately yield

$$\sum_{s=K+1}^{t-1} \delta_s \leqslant \sqrt{\sum_{s=K+1}^{t} \sum_{a=1}^{K} q_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 \ln K} + \frac{(M-m)K}{\gamma_t}\left(1 + \frac{2}{3}\ln K\right)$$
$$\leqslant \sqrt{2}\sqrt{\sum_{s=K+1}^{t} \sum_{a=1}^{K} p_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 \ln K} + \frac{(M-m)K}{\gamma_t}\left(1 + \frac{2}{3}\ln K\right).$$

Unlike what we did to complete the proof of Theorem 4.3, we do not take expectations and rather proceed with deterministic bounds. By the definition (4.8) of the estimated payoffs for the equality below, by (4.10) for the first inequality below, and by the fact that the exploration rates are non-increasing for the second inequality below, we have, for all $s \geqslant K+1$,

$$\sum_{a=1}^{K} p_{s,a}\bigl(\widehat{y}_{s,a} - C\bigr)^2 = \frac{\bigl(y_{s,A_s} - C\bigr)^2}{p_{s,A_s}} \leqslant \frac{(M-m)^2}{\gamma_s/K} \leqslant \frac{(M-m)^2}{\gamma_t/K}\,. \tag{4.19}$$

Therefore,

$$\sum_{s=K+1}^{t-1} \delta_s \leqslant \sqrt{2}(M-m)\sqrt{\frac{t\,K\ln K}{\gamma_t}} + \frac{(M-m)K}{\gamma_t}\left(1 + \frac{2}{3}\ln K\right) \stackrel{\text{def}}{=} D_t = \Theta\Bigl(\sqrt{t/\gamma_t} + 1/\gamma_t\Bigr)\,.$$

For the sake of concision, we denoted by D_t the obtained bound. Via the definition (4.18) of η_t, the sum of interest is in turn bounded by

$$\sum_{t=K+1}^{T} \exp\left(-\eta_t\bigl(t-1-K\bigr)\frac{\Delta_a}{2}\right) \leqslant \sum_{t=K+1}^{T} \exp\left(-\frac{\Delta_a \ln K}{2}\;\frac{t-1-K}{D_t}\right) = \mathcal{O}(1)\,,$$

where the equality to $\mathcal{O}(1)$, i.e., the fact that the considered series is bounded, follows from the fact that

$$-(t-1-K)/D_t = \Theta\Bigl(\sqrt{t\gamma_t} + t\gamma_t\Bigr) = \Theta\bigl(t^{(1-\alpha)/2} + t^{1-\alpha}\bigr)\,.$$

Second sum in the right-hand side of (4.17). We will use Bernstein's inequality for martingales, and more specifically, the formulation of the inequality by Freedman [1975, Thm. 1.6] (see also Massart [2007, Section 2.2]), as stated next.

Reminder 4.3. *Let $(X_n)_{n\geqslant 1}$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_n)_{n\geqslant 0}$, and let $N \geqslant 1$ be a summation horizon. Assume that there exist real numbers b and v_N such that, almost surely,*

$$\forall n \leqslant N, \quad X_n \leqslant b \qquad \text{and} \qquad \sum_{n=1}^{N} \mathbb{E}\bigl[X_n^2 \,\big|\, \mathcal{F}_{n-1}\bigr] \leqslant v_N\,.$$

Then for all $\delta \in (0,1)$,

$$\mathbb{P}\left[\sum_{n=1}^{N} X_n \geqslant \sqrt{2v_N \ln\frac{1}{\delta}} + \frac{b}{3}\ln\frac{1}{\delta}\right] \leqslant \delta\,.$$

For $s \geqslant K+1$, we consider the increments $X_s = \Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}$, which are adapted to the filtration $\mathcal{F}_s = \sigma(A_1, Z_1, \ldots, A_s, Z_s)$, where we recall that $Z_1, \ldots, Z_s$ denote the payoffs obtained in rounds $1, \ldots, s$. Also, as p_s is measurable with respect to past information $\mathcal{F}_{s-1}$ and

since payoffs are drawn independently from everything else (see Section 4.2), we have, by the definition (4.8) of the estimated payoffs (where we rather denote by $Y_{s,a}$ the payoffs drawn at random according to ν_a, to be in line with the notation of Section 4.2 for stochastic bandits): for all $a \in [K]$,

$$\mathbb{E}\big[\,\widehat{y}_{s,a}\,\big|\,\mathcal{F}_{s-1}\big] = \frac{\mathbb{E}[Y_{s,a}\,|\,\mathcal{F}_{s-1}] - C}{p_{s,a}}\mathbb{1}_{\{A_s=a\}} + C = \frac{\mu_a - C}{p_{s,a}}\mathbb{1}_{\{A_s=a\}} + C = \mu_a\,.$$

As a consequence, $\mathbb{E}[X_s\,|\,\mathcal{F}_{s-1}] = \mathbb{E}\big[\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\,|\,\mathcal{F}_{s-1}\big] = 0$. Put differently, $(X_s)_{s\geqslant K+1}$ is indeed a martingale difference sequence with respect to the filtration $(\mathcal{F}_s)_{s\geqslant K}$.

We now check that the additional assumptions of Reminder are satisfied. Manipulations and arguments similar to the ones used in (4.10) and (4.19) show that for all $s \geqslant K+1$,

$$\begin{aligned}\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a} &\leqslant \Delta_a - \frac{Y_{s,a^\star} - C}{p_{s,a}}\mathbb{1}_{\{A_s=a^\star\}} + \frac{Y_{s,a} - C}{p_{s,a}}\mathbb{1}_{\{A_s=a\}} \\ &\leqslant (M-m)(1+K/\gamma_s) \leqslant b \overset{\text{def}}{=} (M-m)(1+K/\gamma_t)\,.\end{aligned}$$

For the variance bound, we first note that for all $s \leqslant t-1$, we have $(\widehat{y}_{s,a} - C)(\widehat{y}_{s,a^\star} - C) = 0$ because of the indicator functions, and therefore,

$$\begin{aligned}\mathbb{E}\Big[\big(\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big] &\leqslant \mathbb{E}\Big[\big(\widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big] \\ &\leqslant \mathbb{E}\Big[\big(\widehat{y}_{s,a^\star} - C\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big] + \mathbb{E}\Big[\big(\widehat{y}_{s,a} - C\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big]\,;\end{aligned}$$

in addition, for all $a \in [K]$ (including $a^\star$),

$$\mathbb{E}\Big[\big(\widehat{y}_{s,a} - C\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big] = \mathbb{E}\left[\frac{(Y_{s,A_s} - C)^2}{p_{s,a}^2}\mathbb{1}_{\{A_s=a\}}\,\middle|\,\mathcal{F}_{s-1}\right] \leqslant \frac{(M-m)^2}{p_{s,a}} \leqslant \frac{(M-m)^2 K}{\gamma_t}\,.$$

Therefore

$$\sum_{s=K+1}^{t-1}\mathbb{E}\Big[\big(\Delta_a - \widehat{y}_{s,a^\star} + \widehat{y}_{s,a}\big)^2\,\Big|\,\mathcal{F}_{s-1}\Big] \leqslant \frac{2K(M-m)^2(t-1-K)}{\gamma_t} \leqslant v_t \overset{\text{def}}{=} \frac{2(M-m)^2 tK}{\gamma_t}\,.$$

Bernstein's inequality (Reminder 4.3) may thus be applied; the choice $\delta = 1/t$ therein leads to

$$\mathbb{P}\left[\sum_{s=K+1}^{t-1}\big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant \underbrace{2(M-m)\sqrt{\frac{tK}{\gamma_t}\ln t} + \frac{M-m}{3}\left(1 + \frac{K}{\gamma_t}\right)\ln t}_{\overset{\text{def}}{=} D'_t}\right] \leqslant \frac{1}{t}\,.$$

As $\sqrt{t/\gamma_t} = \mathcal{O}(t^{(1+\alpha)/2})$ and $1/\gamma_t = \mathcal{O}(t^\alpha)$ as $t \to \infty$, where $\alpha < 1$, and as $\Delta_a > 0$ (given that we are considering a suboptimal arm a), there exists $t_0 \in \mathbb{N}$ such that for all $t \geqslant t_0$,

$$D'_t \leqslant \frac{(t-1-K)\Delta_a}{2}$$

thus

$$\begin{aligned}\mathbb{P}\left[\sum_{s=K+1}^{t-1}(\widehat{y}_{s,a} - \widehat{y}_{s,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2}\right] &= \mathbb{P}\left[\sum_{s=K+1}^{t-1}\big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant \frac{(t-1-K)\Delta_a}{2}\right] \\ &\leqslant \mathbb{P}\left[\sum_{s=K+1}^{t-1}\big(\Delta_a - (\widehat{y}_{s,a^\star} - \widehat{y}_{s,a})\big) \geqslant D'_t\right] \leqslant \frac{1}{t}\,.\end{aligned}$$

Therefore, as $T \to \infty$

$$\sum_{t=1}^{T} \mathbb{P}\left[\sum_{t=K+1}^{t-1} (\widehat{y}_{t,a} - \widehat{y}_{t,a^\star}) \geqslant -\frac{(t-1-K)\Delta_a}{2} \right] = \mathcal{O}(\ln T)\,,$$

as claimed. This concludes the proof. □

4.B. The case of one known end of the payoff range (bandits with gains or with losses)

In this section, we only discuss distribution-free and distribution-dependent upper bounds on the regret, as well as distribution-dependent lower bounds on the regret. This is because the $(M-m)\sqrt{KT}$ distribution-free regret lower bound of Auer et al. [2002b] holds even in the case when both ends m and M of the range are known.

We identified two difficulties in this chapter when the range of bounded payoffs is unknown. First, no $\ln T$ rate for distribution-dependent bounds may be achieved, see (4.4) and Theorem 4.1. Second, there exists a trade-off between distribution-free and distribution-dependent rates for range adaptation, see Theorem 4.2. It turns out that when the upper end M on the payoff range is known, these difficulties (should) disappear. On the contrary, they remain when only the lower end m on the payoff range is known. These statements are detailed and proved below. We therefore contribute to enlightening the difference in nature between bandits with gains and bandits with losses, a topic that was already discussed by Cesa-Bianchi and Lugosi [2006, Remark 6.5, page 164] and Kwon and Perchet [2016].

4.B.1. Known lower end m but unknown upper end M on the payoff range

This case corresponds to considering the model $\mathcal{D}_{m,+}$ defined in (4.5) as

$$\mathcal{D}_{m,+} = \bigcup_{\substack{M\in\mathbb{R},\\ m<M}} \mathcal{D}_{m,M}\,.$$

What is discussed below actually also holds for the larger model $\mathcal{D}_{m,+\infty}$ consisting of probability distributions with a first moment supported $[m,+\infty)$. Note that we have the strict inclusion $\mathcal{D}_{m,+} \subset \mathcal{D}_{m,+\infty}$ as distributions in $\mathcal{D}_{m,+\infty}$ are not bounded in general.

Definitions 4.1 and 4.2 handle the case of $\mathcal{D}_{-,+}$ but can be adapted in an obvious way to $\mathcal{D}_{m,+}$ by fixing m, by having the strategy know m, and require the bounds to hold for all $M \in [m,+\infty)$ and all bandit problems in $\mathcal{D}_{m,M}$. We then refer to scale-free distribution-free regret bounds and distribution-dependent rates for adaptation to the upper end of the range.

We already explained that the construction used to prove Theorem 4.1 not only works for $\mathcal{D}_{-,+}$ but also for $\mathcal{D}_{m,+}$. It turns out that the exact same construction was considered in Theorem 4.2: defining $\nu'_a = (1-\varepsilon)\nu_a + \varepsilon\delta_{\mu_a+2\Delta_a/\varepsilon}$ from a distribution ν_a. When $\nu_a \in \mathcal{D}_{m,+}$, we also have that ν'_a is a bounded distribution, with support lower bounded by m, that is $\nu'_a \in \mathcal{D}_{m,+}$. The proof and thus the result of Theorem 4.2 thus also holds for the case of $\mathcal{D}_{m,+}$.

4.B.2. Known upper end M but unknown lower end m on the payoff range

When the upper end M of the payoff range is known, $\ln T$ distribution-dependent regret rates are possible and there exists an algorithm achieving the optimal problem-dependent constant (Section 4.B.2). Also, $\sqrt{KT}$ scale-free distribution-free regret upper bounds may be achieved (Section 4.B.2), which exactly match the distribution-free lower bound. We could not exhibit a strategy that would simultaneously achieve both optimal distribution-dependent and distribution-free regret bounds, unlike what is known in the case of a known payoff range (the KL-UCB-switch strategy by Garivier et al. [2018]). We however conjecture that this should be possible and that, at least, no trade-off exists between the two bounds (unlike the one imposed by Theorem 4.2).

The case considered in this subsection corresponds to the models $\mathcal{D}_{-,M}$, for $M \in \mathbb{R}$, defined as

$$\mathcal{D}_{-,M} = \bigcup_{\substack{m\in\mathbb{R},\\ m<M}} \mathcal{D}_{m,M}\,.$$

Some of the results actually also hold more generally for semi-bounded payoffs, which correspond to the models $\mathcal{D}_{-\infty,M}$, for $M \in \mathbb{R}$, defined as the sets of probability distributions with a first moment supported on $(-\infty, M]$. Note that we have the strict inclusion $\mathcal{D}_{-,M} \subset \mathcal{D}_{-\infty,M}$ as distributions in $\mathcal{D}_{-\infty,M}$ are not bounded in general.

Known M but unknown m, part 1: distribution-dependent bounds

We may again adapt Definitions 4.1 and 4.2 to define the concepts of distribution-free and distribution-dependent rates for adaptation to the lower end of the range, by considering the models $\mathcal{D}_{-,M}$ or $\mathcal{D}_{-\infty,M}$ therein. The DMED strategy of Honda and Takemura [2015] does achieve a $\ln T$ distribution-dependent rate for adaptation to the lower end of the range and is even competitive against all bandit problems in $\mathcal{D}_{-\infty,M}$. The achieved upper bound is asymptotically optimal as indicated by Reminder 4.1.

Reminder 4.4 (Honda and Takemura [2015], main theorem). *The regret of the DMED strategy is bounded, for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{-\infty,M}$, as*

$$\limsup_{T\to\infty} \frac{R_T(\underline{\nu})}{\ln T} \leqslant \sum_{a=1}^{K} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{-\infty,M})} \,.$$

The nice and deep result of Reminder 4.4 implies that from the distribution-dependent point of view, adaptation to the lower end m of the range is automatic (if such a lower end exists: result holds also when there is no lower bound on the payoffs). Our intuition and understanding for this situation is the following. When the model is $\mathcal{D}_{m,M}$ for known ends m and M, the optimal constant for the $\ln T$ regret is given (see again Reminder 4.1) for all bandit problems $\underline{\nu}$ in $\mathcal{D}_{m,M}$ by

$$C(\underline{\nu}, m, M) = \sum_{a=1}^{K} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}_{m,M})} \,.$$

But it actually turns out, as indicated by Proposition 4.1 below, that $C(\underline{\nu}, m, M)$ is independent of m and equals $C(\underline{\nu}, -\infty, M)$.

Proposition 4.1. *Fix $M \in \mathbb{R}$. For all $m \leqslant M$, for all $\nu \in \mathcal{D}_{m,M}$ and all $\mu > \mathrm{E}(\nu)$,*

$$\mathcal{K}_{\inf}\big(\nu, \mu, \mathcal{D}_{m,M}\big) = \mathcal{K}_{\inf}\big(\nu, \mu, \mathcal{D}_{-\infty,M}\big) \,.$$

Proof. The inequality $\geqslant$ is immediate, as the right-hand side of the equality is an infimum over the larger set $\mathcal{D}_{-\infty,M}$. For the inequality $\leqslant$, we may assume with no loss of generality that $\mu < M$, as otherwise, there is no distribution ν' neither in $\mathcal{D}_{m,M}$ nor in $\mathcal{D}_{-\infty,M}$ with $\mathrm{E}(\nu') > \mu \geqslant M$, so that both $\mathcal{K}_{\inf}$ quantities equal $+\infty$.

We fix M, m, ν and μ as in the statement of the proposition. It suffices to show that in the case $\mu < M$, for all $\nu' \in \mathcal{D}_{-\infty,M}$ with $\mathrm{E}(\nu') > \mu$ and $\nu \ll \nu'$, there exists $\nu'' \in \mathcal{D}_{m,M}$ with $\mathrm{E}(\nu'') > \mu$ and $\mathrm{KL}(\nu, \nu'') \leqslant \mathrm{KL}(\nu, \nu')$. (If ν is not absolutely continuous with respect to ν', then $\mathrm{KL}(\nu, \nu') = +\infty$ and taking ν'' as the Dirac mass δ_M at M is a suitable choice.) To do so, given such a distribution ν', we first note that $\nu \ll \nu'$ and $\nu \in \mathcal{D}_{m,M}$, i.e., $\nu([m, M]) = 1$, entail that $\nu'([m, M]) > 0$, so that we may define the restriction $\nu'' = \nu'_{[m,M]}$ of ν' to $[m, M]$; its density with respect to ν' is given by

$$\frac{\mathrm{d}\nu''}{\mathrm{d}\nu'}(x) = \nu'\big([m, M]\big)^{-1} \mathbb{1}_{\{x\in[m,M]\}} \qquad \nu'\text{–a.s. for all } x \in \mathbb{R}.$$

We have the absolute-continuity chain $\nu \ll \nu'' \ll \nu'$, and the Radon-Nykodym derivatives thus defined satisfy

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(x) = \frac{\mathrm{d}\nu}{\mathrm{d}\nu''}(x)\frac{\mathrm{d}\nu''}{\mathrm{d}\nu'}(x) = \nu'\big([m,M]\big)^{-1}\,\frac{\mathrm{d}\nu}{\mathrm{d}\nu''}(x)\,\mathbb{1}_{\{x\in[m,M]\}} \qquad \nu'\text{–a.s. for all } x\in\mathbb{R}. \quad (4.20)$$

Moreover $\mathrm{E}(\nu'') \geqslant \mathrm{E}(\nu')$, and thus $\mathrm{E}(\nu'') > \mu$, as

$$\begin{aligned}\mathrm{E}(\nu') &= \int_{(-\infty,m)} x\,\mathrm{d}\nu'(x) + \int_{[m,M]} x\,\mathrm{d}\nu'(x)\\ &\leqslant \Big(1-\nu'\big([m,M]\big)\Big)m + \nu'\big([m,M]\big)\ \mathrm{E}(\nu'') \leqslant \mathrm{E}(\nu'')\,.\end{aligned}$$

Finally, by (4.20), which also holds ν–almost surely, and the definition of Kullback-Leibler divergences,

$$\begin{aligned}\mathrm{KL}(\nu,\nu') = \int_{(-\infty,M]} \ln\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)\mathrm{d}\nu &= -\ln\nu'\big([m,M]\big) + \int_{[m,M]} \ln\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu''}\right)\mathrm{d}\nu\\ &= -\ln\nu'\big([m,M]\big) + \mathrm{KL}(\nu,\nu'') \geqslant \mathrm{KL}(\nu,\nu'')\,.\end{aligned}$$

This concludes the proof. □

Known M but unknown m, part 2: distribution-free bounds

A first observation is that (as in the case of a fully known payoff range) AdaHedge does not require any extra-exploration (i.e., any mixing with the uniform distribution) to achieve a scale-free distribution-free regret bound of order $(M-m)\sqrt{KT\ln K}$. This is formally detailed in Appendix 4.C.3. Both this result and the one described next rely on the AdaFTRL methodology of Orabona and Pál [2018], which we recall in Appendix 4.C.1.

The INF strategy of Audibert and Bubeck [2009] can be seen as an instance of FTRL with 1/2–Tsallis entropy, as essentially noted by Audibert et al. [2014]. The INF strategy provides a distribution-free regret bound of order $\sqrt{KT}$ in case of a known payoff range. Up to some technical issues, which we could solve, it may be extended to provide a similar scale-free distribution regret bound, which is optimal as it does not contain any superfluous $\sqrt{\ln K}$ factor. The exact statement to be proved in Appendix 4.C.4 is the following: AdaFTRL with 1/2–Tsallis entropy relying on an upper bound M on the payoffs ensures that for all $m\in\mathbb{R}$ with $m\leqslant M$, for all oblivious individual sequences $y_1, y_2, \ldots$ in $[m,M]^K$, for all $T\geqslant 1$,

$$R_T(y_{1:T}) \leqslant 4(M-m)\sqrt{KT} + 2(M-m)\,.$$

We conclude this section by providing a high-level idea of the technical issues that were solved to obtain the latter bound. We consider estimates $\widehat{y}_{t,a}$ obtained from (4.8) by replacing the constant C therein by the known upper end M. We however could not simply derive the regret bound from some generic full-information regret guarantee for AdaFTRL with 1/2–Tsallis entropy, as to the best of our knowledge, there are no meaningful full-information regret bounds for Tsallis entropy in the first place, and as these would anyway scale with the effective range of the estimates. We instead provide a more careful analysis exploiting special properties of the estimates: $\widehat{y}_{t,a} = M$ for all $a \neq A_t$ and $\widehat{y}_{t,A_t} \leqslant M$.

Open problem 1. We however were unable so far to provide a non-trivial scale-free distribution-dependent regret bound for our strategy AdaFTRL with 1/2–Tsallis entropy. Note that there exist $\mathcal{O}(\ln T)$ bounds for FTRL with 1/2–Tsallis entropy, i.e., with a different tuning of the

learning rates (namely, η_t of order $1/\sqrt{t}$, but then, the range adaptive distribution-free guarantees are lost); see Zimmert and Seldin [2019]. We would have liked to prove such a $\mathcal{O}(\ln T)$ scale-free distribution-dependent regret bound for AdaFTRL with 1/2–Tsallis entropy (or even achieve a more modest aim like a poly-logarithmic bound), as this seems possible and would have shown with certainty that the trade-off imposed by Theorem 4.2 does not hold anymore when the upper end M on the payoff range is known. The techniques of Seldin and Lugosi [2017], which consist in a precise tuning of the extra-exploration in their variant of the Exp3 algorithm of Auer et al. [2002b] together with a gap estimation scheme, or the ones of Zimmert and Seldin [2019] might be helpful to that end. We leave this problem for future research.

Open problem 2. For the sake of completeness, we underline here that either getting rid of the $\sqrt{\ln K}$ factor in the scale-free distribution-free regret bound of AdaHedge for K–armed bandits in the general case of an unknown upper end M on the payoff range, or, alternatively, exhibiting a larger lower bound of order $\sqrt{KT \ln K}$ for this scale-free distribution-free regret, is also a problem that we could not solve yet.

4.C. Known results on AdaFTRL and AdaHedge in full information and applications thereof in the bandit setting

The aim of this section is two-fold: first, we provide, for the sake of self-completeness, a proof of the full-information bound for AdaHedge (Reminder 4.2 in Appendix 4.A); second, we state and prove the improved bandit regret bounds alluded at in Appendix 4.B.2, in the case of a known upper end M but unknown lower end m of the payoff range. We do respectively so in Appendices 4.C.2 (for the full-information bound for AdaHedge) and in Appendices 4.C.3 and 4.C.4 (for the improved bandit regret bounds).

All these bounds can be put under the umbrella of the AdaFTRL methodology of Orabona and Pál [2018] (AdaFTRL stands for adaptive follow-the-regularized-leader), which we recall, again for the sake of self-completeness, in Section 4.C.1. This AdaFTRL methodology was partially built on and inspired the analysis for AdaHedge, which is a special case of AdaFTRL with entropic regularizer (see De Rooij et al. [2014] for AdaHedge, as well as the earlier analysis by Cesa-Bianchi et al. [2007]). Koolen [2016] actually proposes an alternative analysis of AdaFTRL, closer to the AdaHedge formulation, namely, using directly some mixability gaps instead of upper bounds thereon; this is the analysis we recall below in Section 4.C.1.

4.C.1. AdaFTRL for full information (reminder of known results)

To avoid confusion with the notation used in the main body of the chapter, we first describe the considered setting of prediction of oblivious individual sequences with full information.

Full-information setting. The game between the player and the environment is actually the same as the one described in Section 4.2.2, except that the player observes at each step the entire payoff vector, not just the obtained payoff. More formally (and with a different piece of notation z instead of y, to better distinguish the two settings), the environment first picks a sequence of payoff vectors $z_t \in \mathbb{R}^K$, for all $t \geqslant 1$. Then, in a sequential manner, at every time step t, the player picks an action A_t, distributed according to a probability p_t over the action set $[K]$, obtains the payoff z_{t,A_t}, and observes the entire vector z_t (i.e., also the payoffs $z_{t,a}$ corresponding to the actions $a \neq A_t$).

In the sequel, we denote by $\mathcal{S}$ the simplex of probability distributions over $[K]$ and we use the short-hand notation, for $p \in \mathcal{S}$ and $z \in \mathbb{R}^K$,

$$\langle p,\, z\rangle = \sum_{a\in[K]} p_a z_a \,.$$

FTRL (follow-the-regularized-leader). The FTRL method consists in choosing p_t according to

$$p_t \in \operatorname*{argmin}_{p\in\mathcal{S}:F(p)<+\infty} \left\{ \frac{F(p)}{\eta_t} - \sum_{s=1}^{t-1} \langle p,\, z_s\rangle \right\},$$

where $F : \mathbb{R}^K \to \mathbb{R}\cup\{+\infty\}$ is a convex function, called the regularizer, and η_t is a non-negative learning rate in $(0, +\infty]$, which may depend on past observations. The condition $F(p) < +\infty$ will always be satisfied for some $p \in \mathcal{S}$ by the considered regularizers (see below) and is only meant to avoid the undefined $+\infty/+\infty$ in the case $\eta_t = +\infty$. For the sake of concision we will however omit it in the sequel.

Let us give a succint account of the convex analysis results we use here, following the exposition of Lattimore and Szepesvári [2020, Chapter 26]. Using their terminology, the domain $\operatorname{Dom} L$ of a convex function $L : \mathbb{R}^K \to \mathbb{R}\cup\{+\infty\}$ is the set $\{x \in \mathbb{R}^K : L(x) < +\infty\}$ of those points where it

takes finite values. A convex function $L : \mathbb{R}^K \to \mathbb{R} \cup \{+\infty\}$ is said to be Legendre if the interior of its domain $\mathrm{Int}(\mathrm{Dom}\, L)$ is non-empty, if L is strictly convex and differentiable on $\mathrm{Int}(\mathrm{Dom}\, L)$, and if its gradient ∇L blows up on the boundary of $\mathrm{Dom}\, L$. The minimizers of Legendre functions may be seen to satisfy the following properties.

Proposition 4.2 (Special case of Lattimore and Szepesvári [2020, Proposition 26.14])**.** *Let L be a Legendre function and $A \subseteq \mathbb{R}^d$ be a convex set that intersects $\mathrm{Int}(\mathrm{Dom}\, L)$. Then L possesses a unique minimizer $x^\star$ over A, which belongs to $\mathrm{Int}(\mathrm{Dom}\, L)$, therefore ensuring that L is differentiable at $x^\star$. Furthermore,*

$$\forall x \in A \cap \mathrm{Dom}\, L, \qquad \langle \nabla L(x^\star),\, x - x^\star \rangle \geqslant 0\,.$$

Finally, for $x, y \in \mathbb{R}^d$, if $F : \mathbb{R}^K \to \mathbb{R} \cup \{+\infty\}$ is differentiable at y, we define the Bregman divergence between x and y as

$$B_F(x, y) = F(x) - F(y) - \langle \nabla F(y),\, x - y \rangle\ ; \tag{4.21}$$

when F is convex, we have $B_F(x, y) \geqslant 0$ for all $x \in \mathbb{R}^d$.

We are now ready to state our first reminder, which is a classical regret bound for FTRL (see, e.g., Lattimore and Szepesvári [2020, Chapter 28, Exercise 28.12] for references, and McMahan [2017] for more general versions). It involves the diameter D_F of the action set (the K–dimensional simplex $\mathcal{S}$ in our case):

$$D_F = \max_{p,q \in \mathcal{S}} \big\{ F(p) - F(q) \big\}\,.$$

Reminder 4.5 (Generic full-information FTRL bound over the simplex)**.** *The FTRL method with a Legendre regularizer F (of finite diameter D_F) and with any rule for picking the learning rates so that they form a non-increasing sequence satisfies the following guarantee: for all sequences $z_1, z_2, \ldots$ of vector payoffs in $\mathbb{R}^K$, the regret is bounded by*

$$\begin{aligned} \max_{a \in [K]} \sum_{t=1}^{T} z_{t,a} - \sum_{t=1}^{T} \langle p_t,\, z_t \rangle \leqslant \frac{D_F}{\eta_T} &+ \sum_{t=1}^{T-1} \left(\langle p_t - p_{t+1},\, -z_t \rangle - \frac{B_F(p_{t+1}, p_t)}{\eta_t} \right) \\ &+ \left(\langle p_T - p^\star,\, -z_T \rangle - \frac{B_F(p^\star, p_T)}{\eta_T} \right), \end{aligned} \tag{4.22}$$

$$\text{where} \qquad p^\star \in \operatorname*{argmax}_{p \in \mathcal{S}} \sum_{t=1}^{T} \langle p,\, z_t \rangle$$

and where the regret bound is well defined, thanks to the following observations and conventions: for rounds $t \geqslant 1$ where $\eta_t < +\infty$, the function F is indeed differentiable at p_t so that $B_F(p_{t+1}, p_t)$ is well defined; for rounds $t \geqslant 1$ where $\eta_t = +\infty$, we set $B_F(p_{t+1}, p_t)/\eta_t = 0$ irrespectively of the fact whether F is differentiable at p_t.

Proof of Reminder 4.5. Denote by S_t the cumulative vector payoff up to time $t \geqslant 1$. Fix $T \geqslant 1$. For the sake of concision of the equations, we define $p_{T+1} = p^\star$, which is a Dirac mass at some arm (that is, p_{T+1} is not given by FTRL). The regret can therefore be rewritten as

$$\begin{aligned} \max_{a \in [K]} \sum_{t=1}^{T} z_{t,a} - \sum_{t=1}^{T} \langle p_t,\, z_t \rangle &= \max_{p \in \mathcal{S}} \sum_{t=1}^{T} \langle p,\, z_t \rangle - \sum_{t=1}^{T} \langle p_t,\, z_t \rangle \\ &= \sum_{t=1}^{T} \langle p_{T+1},\, z_t \rangle - \sum_{t=1}^{T} \langle p_t,\, z_t \rangle = \sum_{t=1}^{T} \langle p_t - p_{T+1},\, -z_t \rangle\ . \end{aligned}$$

By summation by parts,

$$\sum_{t=1}^{T} \langle p_t - p_{T+1},\, -z_t\rangle = \sum_{t=1}^{T}\sum_{s=t}^{T} \langle p_s - p_{s+1},\, -z_t\rangle = \sum_{s=1}^{T}\sum_{t=1}^{s} \langle p_s - p_{s+1},\, -z_t\rangle = \sum_{s=1}^{T} \langle p_s - p_{s+1},\, -S_s\rangle$$
$$= \sum_{t=1}^{T} \langle p_t - p_{t+1},\, -z_t\rangle + \sum_{t=1}^{T} \langle p_t - p_{t+1},\, -S_{t-1}\rangle \,. \tag{4.23}$$

If $\eta_t < +\infty$, then by the optimality condition from Proposition 4.2 applied to the Legendre function $L : x \mapsto \eta_t^{-1}F(x) - \langle S_{t-1},\, x\rangle$, we know that L thus F are differentiable at p_t and that

$$\big\langle \eta_t^{-1}\nabla F(p_t) - S_{t-1},\, p_{t+1} - p_t\big\rangle \geqslant 0\,,$$
$$\text{that is,} \qquad \langle p_t - p_{t+1},\, -S_{t-1}\rangle \leqslant \big\langle \eta_t^{-1}\nabla F(p_t),\, p_{t+1} - p_t\big\rangle \,.$$

If $\eta_t = +\infty$, the previous inequality holds too, as by definition of p_t, we have $\langle p_t - p_{t+1},\, -S_{t-1}\rangle \leqslant 0$ and as we set by convention $\eta_t^{-1}\nabla F(p_t) = 0$ regardless of whether F is differentiable at p_t or not. Substituting in (4.23), we proved so far

$$\sum_{t=1}^{T} \langle p_t - p_{T+1},\, -z_t\rangle \leqslant \sum_{t=1}^{T} \langle p_t - p_{t+1},\, -z_t\rangle + \big\langle \eta_t^{-1}\nabla F(p_t),\, p_{t+1} - p_t\big\rangle \,. \tag{4.24}$$

This inequality can be rewritten in terms of Bregman divergences:

$$\sum_{t=1}^{T} \langle p_t - p^\star,\, -z_t\rangle \leqslant \sum_{t=1}^{T} \left(\langle p_t - p_{t+1},\, -z_t\rangle - \frac{B_F(p_{t+1}, p_t)}{\eta_t}\right) + \sum_{t=1}^{T} \frac{F(p_{t+1}) - F(p_t)}{\eta_t}$$

We now upper bound the second sum in the right-hand side: again by summation by parts, with the convention $\eta_0 = +\infty$ and $1/\eta_0 = 0$:

$$\sum_{t=1}^{T} \frac{F(p_{t+1}) - F(p_t)}{\eta_t} = \sum_{t=1}^{T} \big(F(p_{t+1}) - F(p_t)\big) \sum_{s=1}^{t} \left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right)$$
$$= \sum_{s=1}^{T}\sum_{t=s}^{T} \big(F(p_{t+1}) - F(p_t)\big) \left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right) = \sum_{s=1}^{T} \big(\underbrace{F(p_{T+1}) - F(p_s)}_{\leqslant D_F}\big) \underbrace{\left(\frac{1}{\eta_s} - \frac{1}{\eta_{s-1}}\right)}_{\geqslant 0} \leqslant \frac{D_F}{\eta_T}\,,$$

where the final equality is obtained by a telescoping sum, using that the sequence of learning rates is non-increasing. □

AdaFTRL, an adaptive version of FTRL. The AdaFTRL approach consists in tuning the learning rate in a way that scales with the observed data. More precisely, it relies on a quantity called the (generalized) mixability gap, which naturally appears as an upper bound on the summands in the FTRL bound of Reminder 4.5:

$$\delta_t^F \overset{\text{def}}{=} \max_{p\in\mathcal{S}} \left\{ \langle p_t - p,\, -z_t\rangle - \frac{B_F(p, p_t)}{\eta_t}\right\} \geqslant 0\,. \tag{4.25}$$

That mixability gaps are always nonnegative can be seen by taking $p = p_t$ in the definition. We may further upper bound (4.22) when it holds by using this mixability gap:

$$\max_{a\in[K]} \sum_{t=1}^{T} z_{t,a} - \sum_{t=1}^{T} \langle p_t,\, z_t\rangle \leqslant \frac{D_F}{\eta_T} + \sum_{t=1}^{T} \delta_T^F\,. \tag{4.26}$$

The AdaFTRL learning rate balances the two terms in the above regret bound by taking

$$\eta_t = D_F \Big/ \sum_{s=1}^{t-1} \delta_s^F \quad \in (0, +\infty] \tag{4.27}$$

Note that this rule for picking learning rates indeed leads to non-increasing sequences thereof, as the mixability gaps are non-negative. We summarize the discussion above in the theorem stated next, from which subsequent (closed-from) regret bounds will be derived by using the specific properties of the regularizer F at hand to upper bound the mixability gaps.

Theorem 4.5 (AdaFTRL tool box). *Under the assumptions of Reminder 4.5 and with its conventions, the regret of the FTRL method based on the learning rates* (4.27) *satisfies*

$$\max_{a\in[K]} \sum_{t=1}^{T} z_{t,a} - \sum_{t=1}^{T} \langle p_t,\, z_t\rangle \leqslant 2 \sum_{t=1}^{T} \delta_t^F \tag{4.28}$$

where, moreover,

$$\left(\sum_{t=1}^{T} \delta_t^F\right)^2 = 2D_F \sum_{t=1}^{T} \frac{\delta_t^F}{\eta_t} + \sum_{t=1}^{T} \big(\delta_t^F\big)^2 . \tag{4.29}$$

Proof. Inequality (4.28) follows from (4.26) and (4.27). The equality (4.29) is obtained by expanding the squared sum,

$$\left(\sum_{t=1}^{T} \delta_t^F\right)^2 = \sum_{t=1}^{T} \big(\delta_t^F\big)^2 + 2\sum_{t=1}^{T}\sum_{s=1}^{t-1} \delta_t^F \delta_s^F = \sum_{t=1}^{T} (\delta_t^F)^2 + 2\sum_{t=1}^{T} \delta_t^F \frac{D_F}{\eta_t}$$

where the final equality is obtained by substituting the definition (4.27) of η_t. □

4.C.2. AdaHedge for full information (reminder of known results)

The content of this section is extracted from various sources, out of which the most important is Koolen [2016]. We claim no novelty. This section recalls how the bound for AdaHedge (Reminder 4.2, for which a direct proof was provided by De Rooij et al. [2014]) can also be seen as a special case of the results of Section 4.C.1.

It is well-known (see Freund et al. [1997], Kivinen and Warmuth [1999], Audibert [2009]) and can be found again by a simple optimization under a linear constraint that the Hedge weight update corresponds to FTRL with the negentropy as a regularizer:

$$H_{\mathrm{neg}}(p) = \sum_{a=1}^{K} p_a \ln p_a \,,$$

with value $+\infty$ whenever $p_a = 0$ for some $a \in [K]$. That is,

$$\operatorname*{argmin}_{p\in\mathcal{S}} \left\{ \frac{H_{\mathrm{neg}}(p)}{\eta_t} - \sum_{s=1}^{t-1} \langle p,\, z_s\rangle \right\} = \{p_t\}$$

$$\text{with} \qquad p_{t,a} = \exp\left(\eta_t \sum_{s=1}^{t-1} z_{a,s}\right) \Big/ \sum_{k=1}^{K} \exp\left(\eta_t \sum_{s=1}^{t-1} z_{k,s}\right). \tag{4.30}$$

Straightforward calculation show that the regularizer H_{neg} is indeed Legendre (see Lattimore and Szepesvári [2020], Example 26.11) and the H_{neg}–diameter of the simplex equals $D_{H_{\text{neg}}} = \ln K$. Reminder 4.5 and Theorem 4.5 can therefore be applied.

AdaHedge is exactly AdaFTRL with H_{neg} as a regularizer. Indeed, the mixability gap (4.25) can be computed in closed form (as noted by Reid et al. [2015, Lemma 5]) and reads in this case:

$$\delta_t^{\text{neg}} = \begin{cases} -\langle p_t,\, z_t\rangle + \eta_t^{-1} \ln\left(\sum_{a=1}^K p_{t,a} \mathrm{e}^{\eta_t z_{t,a}}\right) & \text{if } \eta_t < +\infty, \\ -\langle p_t,\, z_t\rangle + \max_{a\in[K]} z_{t,a} & \text{if } \eta_t = +\infty. \end{cases} \tag{4.31}$$

Proof of the rewriting (4.31). When $\eta_t = +\infty$, the mixability gap equals, by definition,

$$\delta_t^F = \max_{p\in\mathcal{S}} \big\{ \langle p_t - p,\, -z_t\rangle \big\} = -\langle p_t,\, z_t\rangle + \max_{p\in\mathcal{S}} \langle p,\, z_t\rangle = -\langle p_t,\, z_t\rangle + \max_{a\in[K]} z_{t,a}\,.$$

For the case $\eta_t < +\infty$, the following formula, which is at the heart of the closed-form formula for the Hedge updates (4.30), will be useful: for any $S \in \mathbb{R}^d$,

$$\min_{p\in\mathcal{S}} \Big\{ H_{\text{neg}}(p) - \langle p,\, S\rangle \Big\} = \sum_{i=1}^K \frac{\mathrm{e}^{S_i}}{\sum_{j=1}^K \mathrm{e}^{S_j}} \left(\ln\left(\frac{\mathrm{e}^{S_i}}{\sum_{j=1}^K \mathrm{e}^{S_j}}\right) - S_i \right) = -\ln\left(\sum_{i=1}^K \mathrm{e}^{S_i}\right). \tag{4.32}$$

When $\eta_t < +\infty$, Equation (4.30) shows that p_t lies in the interior $\mathrm{Int}(\mathcal{S})$ of $\mathcal{S}$. The Bregman divergence at hand in the definition (4.25) of the mixability gaps may be simplified into

$$B_F(p, p_t) = H_{\text{neg}}(p) - H_{\text{neg}}(p_t) - \langle \nabla H_{\text{neg}}(p_t),\, p - p_t\rangle = H_{\text{neg}}(p) - \langle \nabla H_{\text{neg}}(p_t),\, p\rangle + 1\,,$$

where the second inequality holds by taking into account the fact that H_{neg} is twice differentiable at any $p \in \mathrm{Int}(\mathcal{S})$, with

$$\nabla H_{\text{neg}}(p) = \big(1 + \ln p_i\big)_{i\in[K]} \qquad \text{so that} \qquad \langle \nabla H_{\text{neg}}(p),\, p\rangle = 1 + \sum_{i=1}^K p_i \ln p_i = 1 + H_{\text{neg}}(p)\,.$$

The mixability gaps can therefore be rewritten

$$\begin{aligned} \delta_t^F &= \max_{p\in\mathcal{S}} \left\{ \langle p_t - p,\, -z_t\rangle - \frac{B_F(p, p_t)}{\eta_t} \right\} \\ &= -\langle p_t,\, z_t\rangle - \frac{1}{\eta_t} + \frac{1}{\eta_t} \max_{p\in\mathcal{S}} \big\{ \eta_t \langle p,\, z_t\rangle - H_{\text{neg}}(p) + \langle \nabla H_{\text{neg}}(p_t),\, p\rangle \big\} \\ &= -\langle p_t,\, z_t\rangle - \frac{1}{\eta_t} - \frac{1}{\eta_t} \min_{p\in\mathcal{S}} \Big\{ H_{\text{neg}}(p) - \big\langle p,\, \eta_t z_t + \nabla H_{\text{neg}}(p_t)\big\rangle \Big\} \end{aligned}$$

Now by (4.32), specialized with $S = \eta_t z_t + \nabla H_{\text{neg}}(p_t)$, we can compute the value of the minimum:

$$\min_{p\in\mathcal{S}} \Big\{ H_{\text{neg}}(p) - \big\langle p,\, \eta_t z_t + \nabla H_{\text{neg}}(p_t)\big\rangle \Big\} = -\ln\left(\sum_{i=1}^K \mathrm{e}^{\eta_t z_i + 1 + \ln p_i}\right) = -1 - \ln\left(\sum_{i=1}^K p_i \mathrm{e}^{\eta_t z_i}\right).$$

Collecting all equalities together concludes the proof. □

Reminder 4.2 is thus a special case of the following bound.

Theorem 4.6 (See Lemma 3 and Theorem 6 of De Rooij et al. [2014]). *For all sequences of payoffs $z_{t,a}$ lying in some bounded real-valued interval, denoted by $[b, B]$, for all $T \geqslant 1$, the regret of the AdaHedge algorithm with full information, as defined by (4.30) and (4.31), satisfies*

$$\max_{k\in[K]} \sum_{t=1}^{T} z_{t,k} - \sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\, z_{t,a} \leqslant 2\sum_{t=1}^{T} \delta_t^{neg}$$

$$\text{where} \qquad \sum_{t=1}^{T} \delta_t^{neg} \leqslant \sqrt{\sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\left(z_{t,a} - \sum_{k\in[K]} q_{t,k}\, z_{t,k}\right)^2 \ln K} + (B-b)\left(1 + \frac{2}{3}\ln K\right),$$

and AdaHedge does not require the knowledge of $[b, B]$ to achieve this bound.

The quantities

$$v_t \stackrel{\text{def}}{=} \sum_{a=1}^{K} p_{t,a}\left(z_{t,a} - \sum_{k\in[K]} q_{t,k}\, z_{t,k}\right)^2$$

in the bound correspond to the variance of the random variables taking values $z_{t,a}$ with probability $p_{t,a}$; the variational formula for variances indicates that

$$\sum_{a=1}^{K} p_{t,a}\left(z_{t,a} - \sum_{k\in[K]} q_{t,k}\, z_{t,k}\right)^2 = \min_{c\in\mathbb{R}} \sum_{a=1}^{K} p_{t,a}\big(z_{t,a} - c\big)^2,$$

which entails the final bound given as a note in the statement of Reminder 4.2.

The following formulation of Bernstein's inequality will be useful in the proof of Theorem 4.6.

Lemma 4.1 (Bernstein's inequality tailored to our needs). *Let X be a random variable in $[0, 1]$, with variance denoting by $\mathrm{Var}(X)$. Then for all $\eta > 0$,*

$$\frac{\ln\Big(\mathbb{E}\big[\mathrm{e}^{\eta(X-\mathbb{E}[X])}\big]\Big)}{\eta^2} \leqslant \frac{1}{2}\mathrm{Var}(X) + \frac{1}{3}\,\frac{\ln\Big(\mathbb{E}\big[\mathrm{e}^{\eta(X-\mathbb{E}[X])}\big]\Big)}{\eta}.$$

Proof . Denote by $\psi_X(\eta) = \ln\big(\mathbb{E}\big[\mathrm{e}^{\eta(X-\mathbb{E}[X])}\big]$ the log-moment generating function of X. A version of Bernstein's inequality with an appropriate control of the moments (as stated by Massart [2007, Section 2.2.3] and applied to X with $c = 1/3$) indicates that for all $\eta \in (0, 3)$,

$$\Big(1 - \frac{\eta}{3}\Big)\psi_X(\eta) \leqslant \frac{\eta^2}{2}\mathrm{Var}(X).$$

Actually, this inequality also holds for $\eta \geqslant 3$ as its left-hand side is non-positive while its right-hand side is nonnegative. The claimed result is derived by rearraging the terms

$$\psi_X(\eta) \leqslant \frac{\eta^2}{2}\mathrm{Var}(X) + \frac{\eta}{3}\psi_X(\eta)$$

and by dividing both sides by η^2. □

Proof of Theorem 4.6. We apply Theorem 4.5. To that end, we first bound the mixability gaps. The rewriting (4.31) (and Jensen's inequality) directly shows that $0 \leqslant \delta_t^{\mathrm{neg}} \leqslant B - b$. We may also prove the bound

$$\frac{\delta_t^{\mathrm{neg}}}{\eta_t} \leqslant \frac{v_t}{2} + \frac{1}{3}(B-b)\delta_t^{\mathrm{neg}}. \tag{4.33}$$

It suffices to do so for $\eta_t < +\infty$. Consider the random variable X taking values $(z_{t,a} - b)/(B - b)$ with probability $p_{t,a}$, for $a \in \{1, \dots, K\}$. The mixability gap can be rewritten as

$$\delta_t^{\mathrm{neg}} = \frac{1}{\eta_t} \psi_X\big(\eta_t(B-b)\big)$$

with the notation of the proof of Lemma 4.1. The variance of X equals $v_t/(B-b)^2$. Lemma 4.1 with $\eta = \eta_t(B-b)$ yields

$$\frac{\delta_t^{\mathrm{neg}}}{\eta_t(B-b)^2} \leqslant \frac{v_t}{2(B-b)^2} + \frac{\delta_t^{\mathrm{neg}}}{3(B-b)}\,.$$

from which we obtain (4.33) by rearranging.

From (4.29) and (4.33), we deduce, together with the bound $(\delta_t^{\mathrm{neg}})^2 \leqslant (B-b)\delta_t^{\mathrm{neg}}$, that

$$\left(\sum_{t=1}^T \delta_t^{\mathrm{neg}}\right)^2 \leqslant (\ln K) \sum_{t=1}^T v_t + (B-b)\left(\frac{2}{3}\ln K + 1\right) \sum_{t=1}^T \delta_t^{\mathrm{neg}}\,.$$

Therefore, using the fact that $x^2 \leqslant a + bx$ implies $x \leqslant \sqrt{a} + b$ for all $a, b, x \geqslant 0$,

$$\sum_{t=1}^T \delta_t^{\mathrm{neg}} \leqslant \sqrt{\ln K \sum_{t=1}^T v_t} + (B-b)\left(\frac{2}{3}\ln K + 1\right),$$

which thanks to (4.28) concludes the proof of Theorem 4.6. □

4.C.3. AdaHedge with known upper bound M on the payoffs (application of Section 4.C.2)

We show how to obtain a scale-free distribution-free regret bound of order $(M-m)\sqrt{KT\ln K}$ with no extra-exploration (including no initial exploration) when an upper bound M on the payoffs is given to the player. We consider Algorithm 4.2, where no mixing takes place (unlike in Algorithm 4.1) and where the probability distributions p_t are directly computed via an AdaHedge update (no need for intermediate probabilities q_t). Note also that we use the estimates (4.8) with the choice $C_t = M$, that is,

$$\widehat{y}_{t,a} = \frac{y_{t,a} - M}{p_{t,a}} \mathbb{1}_{\{A_t = a\}} + M\,. \tag{4.34}$$

The following observation is key in the analysis below: $\widehat{y}_{t,a} = M$ for all $a \neq A_t$ and $\widehat{y}_{t,A_t} \leqslant M$. We will also use, as in the proof of Theorem 4.3,

$$\sum_{a=1}^K p_{t,a}\, \widehat{y}_{t,a} = y_{t,A_t}\,.$$

The performance bound for this simpler algorithm is stated next.

Theorem 4.7. *AdaHedge for K–armed bandits relying on an upper bound M on the payoffs (Algorithm 4.2) ensures that for all $m \in \mathbb{R}$ with $m \leqslant M$, for all oblivious individual sequences $y_1, y_2, \dots$ in $[m, M]^K$, for all $T \geqslant 1$,*

$$R_T(y_{1:T}) \leqslant 2(M-m)\sqrt{KT\ln K} + 2(M-m)\,.$$

Algorithm 4.2 AdaHedge for K–armed bandits, when an upper bound on the payoffs is given

Input: an upper bound M on the payoffs
AdaHedge initialization: $\eta_1 = +\infty$ and $p_1 = (1/K, \dots, 1/K)$
for rounds $t = 1, 2, \dots$ **do**
  Draw an arm $A_t \sim p_t$ (independently at random according to the distribution p_t)
  Get and observe the payoff y_{t,A_t}
  Compute the estimates of all payoffs

$$\widehat{y}_{t,a} = \frac{y_{t,a} - M}{p_{t,a}} \mathbb{1}_{\{A_t = a\}} + M$$

  Compute the mixability gap δ_t based on the distribution p_t and on these estimates:

$$\delta_t = \begin{cases} -\sum_{a=1}^K p_{t,a}\, \widehat{y}_{t,a} + \frac{1}{\eta_t} \ln\left(\sum_{a=1}^K p_{t,a} \mathrm{e}^{\eta_t \widehat{y}_{t,a}}\right) & \text{if } \eta_t < +\infty \\ -\sum_{a=1}^K p_{t,a}\, \widehat{y}_{t,a} + \max_{a \in [K]} \widehat{y}_{t,a} & \text{if } \eta_t = +\infty \end{cases}$$

  Compute the learning rate $\eta_{t+1} = \left(\sum_{s=1}^t \delta_s\right)^{-1} \ln K$
  Define p_{t+1} component-wise as

$$p_{t+1,a} = \exp\left(\eta_{t+1} \sum_{s=1}^t \widehat{y}_{a,s}\right) \Big/ \sum_{k=1}^K \exp\left(\eta_{t+1} \sum_{s=1}^t \widehat{y}_{k,s}\right)$$

end for

The main technical difference with respect to the analysis of Algorithm 4.1 is that the mixability gaps are directly bounded by the range $M - m$. We no longer need to artificially control the size of the estimates (which we did via extra-exploration) to get, in turn, a control of the mixability gaps.

Lemma 4.2 (Improved mixability gap bound). *The mixability gaps of AdaHedge for K–armed bandits relying on an upper bound M on the payoffs (Algorithm 4.2) are bounded, for all $m \in \mathbb{R}$ with $m \leqslant M$, for all oblivious individual sequences $y_1, y_2, \dots$ in $[m, M]^K$, for all $t \geqslant 1$, by*

$$0 \leqslant \delta_t \leqslant M - m \qquad \text{and} \qquad \frac{\delta_t}{\eta_t} \leqslant \frac{1}{2} p_{t,A_t}^{-1} (M - y_{t,A_t})^2.$$

Proof. The fact that $\delta_t \geqslant 0$ holds by definition of the gaps and Jensen's inequality. For $\delta_t \leqslant M - m$, the observations after (4.34) indicate that when $\eta_t = +\infty$,

$$\delta_t = -\sum_{a=1}^K p_{t,a}\, \widehat{y}_{t,a} + \max_{a \in [K]} \widehat{y}_{t,a} = M - \widehat{y}_{t,A_t}\,,$$

while for $\eta_t < +\infty$,

$$\begin{aligned}\delta_t &= -y_{t,A_t} + \frac{1}{\eta_t}\ln\Big((1-p_{t,A_t})\mathrm{e}^{\eta_t M} + p_{t,A_t}\mathrm{e}^{\eta_t M}\mathrm{e}^{\eta_t(y_{t,A_t}-M)/p_{t,A_t}}\Big)\\ &\leqslant M - y_{t,A_t} + \frac{1}{\eta_t}\ln\Big((1-p_{t,A_t}) + p_{t,A_t}\underbrace{\mathrm{e}^{\eta_t(y_{t,A_t}-M)/p_{t,A_t}}}_{\leqslant 1}\Big)\,,\end{aligned}$$

which entails $\delta_t \leqslant M - y_{t,A_t} \leqslant M - m$.

Furthermore, in the case $\eta_t < +\infty$, using the inequality $\mathrm{e}^{-x} \leqslant 1 - x + x^2/2$ valid for $x \geqslant 0$, followed by the inequality $\ln(1+u) \leqslant u$, valid for all $u > -1$, we get

$$\delta_t \leqslant M - \widehat{y}_{t,A_t} + \frac{1}{\eta_t}\ln\Big(\underbrace{1 - p_{A_t,t,} + p_{A_t,t}}_{=1}\underbrace{-\eta_t(M - y_{t,A_t}) + \eta_t^2\frac{(M-y_{t,A_t})^2}{2p_{A_t,t}}}_{=u}\Big) \leqslant \eta_t\frac{(M-y_{t,A_t})^2}{2p_{t,A_t}}\,.$$

The second inequality is trivial in case $\eta_t = +\infty$, as $\delta_t/\eta_t = 0$. □

We are now ready to prove Theorem 4.7.

Proof of Theorem 4.7. As indicated in Section 4.C.2, AdaHedge is a special case of AdaFTRL and the bound of Theorem 4.5 is applicable.

Equation (4.29) and Lemma 4.2, which entails in particular that $\delta_t^2 \leqslant (M-m)\delta_t$, yield

$$\Big(\sum_{t=1}^T \delta_t\Big)^2 = 2(\ln K)\sum_{t=1}^T\frac{\delta_t}{\eta_t} + \sum_{t=1}^T(\delta_t)^2 \leqslant (\ln K)\sum_{t=1}^T p_{t,A_t}^{-1}(M - y_{t,A_t})^2 + (M-m)\sum_{t=1}^T\delta_t\,,$$

which, through the fact that $x^2 \leqslant a + bx$ implies $x \leqslant \sqrt{a} + b$ for all $a, b, x \geqslant 0$, leads in turn to

$$\sum_{t=1}^T \delta_t \leqslant \sqrt{\sum_{t=1}^T p_{t,A_t}^{-1}(M - \widehat{y}_{t,A_t})^2\ln K} + (M-m)\,.$$

Therefore, Equation (4.28) guarantees that

$$\max_{k\in[K]}\sum_{t=1}^T \widehat{y}_{t,k} - \sum_{t=1}^T\underbrace{\sum_{a=1}^K p_{t,a}\,\widehat{y}_{t,a}}_{=y_{t,A_t}} \leqslant 2\sqrt{\sum_{t=1}^T p_{t,A_t}^{-1}(M - \widehat{y}_{t,A_t})^2\ln K} + 2(M-m)\,. \tag{4.35}$$

We conclude the proof by integrating the inequality above and using Jensen's inequality, exactly as in the proof of Theorem 4.3. Indeed, Equation (4.13) therein indicates that

$$R_T(y_{1:T}) = \max_{k\in[K]}\sum_{t=1}^T y_{t,k} - \mathbb{E}\left[\sum_{t=1}^T y_{t,A_t}\right] \leqslant \mathbb{E}\left[\max_{k\in[K]}\sum_{t=1}^T\widehat{y}_{t,k} - \sum_{t=1}^T y_{t,A_t}\right]$$

and, by the same manipulations as in (4.15) and in the equation that follows it,

$$\begin{aligned}\mathbb{E}\left[\sqrt{\sum_{t=1}^T p_{t,A_t}^{-1}(M - \widehat{y}_{t,A_t})^2\ln K}\right] &\leqslant \sqrt{\mathbb{E}\left[\sum_{t=1}^T p_{t,A_t}^{-1}(M - y_{t,A_t})^2\ln K\right]}\\ &= \sqrt{\mathbb{E}\left[\sum_{t=1}^T\sum_{a=1}^K(M-y_{t,a})^2\ln K\right]} \leqslant (M-m)\sqrt{KT\ln K}\end{aligned}$$

The claimed result is obtained by collecting all bounds together. □

4.C.4. AdaFTRL with Tsallis entropy in the case of a known upper bound M on the payoffs

In this section we describe how the AdaHedge learning rate scheme can be used in the FTRL framework with a different regularizer, namely Tsallis entropy, to improve the scale-free distribution-free regret bound into a bound of optimal order $(M-m)\sqrt{KT}$, i.e., without any superfluous $\sqrt{\ln K}$ factor.

Tsallis entropy. We focus on the (rescaled) 1/2–Tsallis entropy, which is defined by

$$H_{1/2}(p) = -\sum_{a=1}^{K} 2\sqrt{p_a}\,.$$

This regularizer is Legendre over the domain $[0,+\infty)^K$ (see Lattimore and Szepesvári [2020, Example 26.10]). Its diameter equals

$$D_{H_{1/2}} = \max_{p\in\mathcal{S}} H_{1/2}(p) - \min_{q\in\mathcal{S}} H_{1/2}(q) = -2 - \big(-2\sqrt{K}\big) = 2\big(\sqrt{K}-1\big)\,, \tag{4.36}$$

as for all $p\in\mathcal{S}$, we have (by concavity of the square root for the right-most inequality)

$$1 \leqslant \sum_{a=1}^{K} p_a \leqslant \sum_{a=1}^{K} \sqrt{p_a} \leqslant \sqrt{K}\,,$$

where 1 is achieved with $p=(1,0,\dots,0)$ and $\sqrt{K}$ with the uniform distribution.

The function $H_{1/2}$ is differentiable at all $q\in(0,+\infty)^K$, with $\nabla H_{1/2}(q) = \big(-1/\sqrt{q_a}\big)_{a\in[K]}$. The Bregman divergence associated with $H_{1/2}$ equals, for $p,q\in\mathcal{S}$ such that $q_a>0$ for all a:

$$\begin{aligned} B_{H_{1/2}}(p,q) &= -2\sum_{a=1}^{K}\sqrt{p_a} + 2\sum_{a=1}^{K}\sqrt{q_a} + \sum_{a=1}^{K}\frac{1}{\sqrt{q_a}}(p_a-q_a) \\ &= -2\sum_{a=1}^{K}\frac{\sqrt{p_a}-\sqrt{q_a}}{2\sqrt{q_a}}\Big(2\sqrt{q}_a - (\sqrt{p_a}+\sqrt{q_a})\Big) = \sum_{a=1}^{K}\frac{(\sqrt{p_a}-\sqrt{q_a})^2}{\sqrt{q_a}}\,. \end{aligned}$$

AdaFTRL with 1/2–Tsallis entropy. We consider FTRL with the 1/2–Tsallis entropy on the estimated losses (4.34):

$$p_t \in \operatorname*{argmin}_{p\in\mathcal{S}} \left\{ \frac{H_{1/2}(p)}{\eta_t} - \sum_{s=1}^{t-1}\langle p,\,\widehat{y}_s\rangle \right\} = \operatorname*{argmin}_{p\in\mathcal{S}} \left\{ -\frac{1}{\eta_t}\sum_{a=1}^{K} 2\sqrt{p_a} - \sum_{a=1}^{K} p_a \sum_{s=1}^{t-1}\widehat{y}_{s,a} \right\}.$$

FTRL with the 1/2–Tsallis entropy was essentially introduced by Audibert and Bubeck [2009] to get rid of a $\sqrt{\ln K}$ factor in the distribution-free regret bound of K–armed adversarial bandits (with known payoff range). It was later noted by Audibert et al. [2014] that it actually is an instance of mirror descent with Tsallis entropy as a regularizer. More recently, Zimmert and Seldin [2019] showed that this regularizer can obtain quasi-optimal regret bounds for both stochastic and adversarial rewards.

We more precisely consider AdaFTRL with the 1/2–Tsallis, that is, we compute the learning rates η_t based on the mixability gaps (4.25); see Algorithm 4.3. We denote by δ_t^{Ts} the mixability gaps (4.25).

On the implementation. For Tsallis entropy, the optimization problems involved in the computation of the updates p_t and of the mixability gaps δ_t^{Ts} admit a (semi-)explicit formula. Indeed, p_t can be computed thanks to the formula, for all $z \in \mathbb{R}^K$,

$$\operatorname*{argmin}_{p\in\mathcal{S}} \big\{ H_{1/2}(p) - \langle p,\, z\rangle \big\} = \operatorname*{argmax}_{p\in\mathcal{S}} \left\{ \langle p,\, z\rangle + \sum_{a=1}^{K} 2\sqrt{p_a} \right\} = \left(\frac{1}{\big(c(z) - z_a\big)^2} \right)_{a\in K}, \quad (4.37)$$

where $c(z)$ is an implicit normalization constant, such that the vector lies in the simplex $\mathcal{S}$ and $c(z) > z_a$ for all $a \in [K]$. This constant $c(z)$ is in fact the Lagrange multiplier associated with the constraint $p_1 + \ldots + p_K = 1$. See Zimmert and Seldin [2019] for more details on how to compute $c(z)$ efficiently, see also Audibert et al. [2014]. To compute the mixability gap, rewrite

$$\begin{aligned}\delta_t^{\mathrm{Ts}} &= \max_{p\in\mathcal{S}} \left\{ \langle p_t - p,\, -\widehat{y}_t\rangle - \frac{H_{1/2}(p) - H_{1/2}(p_t) - \big\langle \nabla H_{1/2}(p_t),\, p - p_t\big\rangle}{\eta_t} \right\} \\ &= \langle p_t,\, -\widehat{y}_t\rangle + \frac{H_{1/2}(p_t)}{\eta_t} - \frac{\big\langle \nabla H_{1/2}(p_t),\, p_t\big\rangle}{\eta_t} + \frac{1}{\eta_t} \max_{p\in\mathcal{S}} \Big\{ \big\langle p,\, \nabla H_{1/2}(p_t) + \eta_t \widehat{y}_t\big\rangle - H_{1/2}(p) \Big\}, \end{aligned} \quad (4.38)$$

where the maximum in the left-most side of these equalities can be computed efficiently, thanks to (4.37).

Analysis of the algorithm. We provide the following performance bound.

Theorem 4.8. *AdaFTRL with $1/2$–Tsallis entropy for K–armed bandits relying on an upper bound M on the payoffs (Algorithm 4.3) ensures that for all $m \in \mathbb{R}$ with $m \leqslant M$, for all oblivious individual sequences $y_1, y_2, \ldots$ in $[m, M]^K$, for all $T \geqslant 1$,*

$$R_T(y_{1:T}) \leqslant 4(M-m)\sqrt{KT} + 2(M-m)\,.$$

As in Section 4.C.3, the proof scheme is a combination of the AdaFTRL bound of Theorem 4.5 (which is indeed applicable), together with an improved bound on the mixability gap that exploits the specific shape of the estimates. This bound is stated in the next lemma, which is much similar to Lemma 4.2.

Lemma 4.3. *The mixability gaps of AdaFTRL with Tsallis entropy for K–armed bandits relying on an upper bound M on the payoffs (Algorithm 4.3) are bounded, for all $m \in \mathbb{R}$ with $m \leqslant M$, for all oblivious individual sequences $y_1, y_2, \ldots$ in $[m, M]^K$, for all $t \geqslant 1$, by*

$$0 \leqslant \delta_t^{Ts} \leqslant M - m \qquad \text{and} \qquad \frac{\delta_t^{Ts}}{\eta_t} \leqslant p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2\,.$$

The proof of Lemma 4.3 is postponed to the end of this section and we now proceed with the proof of Theorem 4.8.

Proof of Theorem 4.8. The structure of the proof is much similar to the one of Theorem 4.7, which is why we only sketch our arguments. The bound of Theorem 4.5 is applicable. We use Lemma 4.3 with (4.29) to see that

$$\left(\sum_{t=1}^{T} \delta_t^{\mathrm{Ts}} \right)^2 \leqslant 2D_{H_{1/2}} \sum_{t=1}^{T} p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2 + (M-m) \sum_{t=1}^{T} \delta_t^{\mathrm{Ts}}\,. \quad (4.39)$$

Algorithm 4.3 AdaFTRL with Tsallis entropy for K–armed bandits, when an upper bound on the payoffs is given

Input: an upper bound M on the payoffs
Initialization: $\eta_1 = +\infty$ and $p_1 = (1/K, \dots, 1/K)$
for rounds $t = 1, 2, \dots$ **do**
  Draw an arm $A_t \sim p_t$ (independently at random according to the distribution p_t)
  Get and observe the payoff y_{t,A_t}
  Compute the estimates of all payoffs

$$\widehat{y}_{t,a} = \frac{y_{t,a} - M}{p_{t,a}} \mathbb{1}_{\{A_t = a\}} + M$$

  Compute the mixability gap δ_t^{Ts} based on the distribution p_t and on these estimates, e.g., using the efficient implementation stated around (4.38):

$$\delta_t^{\mathrm{Ts}} = \max_{p \in \mathcal{S}} \left\{ \langle p_t - p, -\widehat{y}_t \rangle - \frac{B_{H_{1/2}}(p, p_t)}{\eta_t} \right\}$$

  Compute the learning rate $\eta_{t+1} = 2\left(\sum_{s=1}^{t} \delta_s^{\mathrm{Ts}} \right)^{-1} \left(\sqrt{K} - 1 \right)$
  Define p_{t+1} as

$$p_{t+1} \in \operatorname*{argmin}_{p \in \mathcal{S}} \left\{ - \sum_{a=1}^{K} p_a \sum_{s=1}^{t} \widehat{y}_{s,a} - \frac{1}{\eta_{t+1}} \sum_{a=1}^{K} 2\sqrt{p_a} \right\},$$

  where an efficient implementation is provided by, e.g., (4.37)
end for

Again, using the fact that for all $a, b, x \geqslant 0$, the inequality $x^2 \leqslant a + bx$ implies $x \leqslant \sqrt{a} + b$:

$$\sum_{t=1}^{T} \delta_t^{\mathrm{Ts}} \leqslant \sqrt{2 D_{H_{1/2}} \sum_{t=1}^{T} p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2} + (M - m) \tag{4.40}$$

By (4.28), by taking expectations, and by Jensen's inequality:

$$R_T(y_{1:T}) \leqslant 2\mathbb{E}\Big[\sum_{t=1}^{T} \delta_t^{\mathrm{Ts}} \Big] \leqslant 2\sqrt{2 D_{H_{1/2}} \sum_{t=1}^{T} \mathbb{E}\Big[p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2 \Big]} + 2(M - m)\,. \tag{4.41}$$

We conclude by observing that for all t, by definition of the payoff estimates,

$$\begin{aligned} \mathbb{E}\Big[p_{t,A_t}^{-1/2} \big(M - y_{t,A_t}\big)^2 \Big] = \mathbb{E}\left[\sum_{a=1}^{K} p_{t,a}\, p_{t,a}^{-1/2} \big(M - y_{t,a}\big)^2 \right] &\leqslant (M - m)^2\, \mathbb{E}\left[\sum_{a=1}^{K} \sqrt{p_{a,t}} \right] \\ &\leqslant (M - m)^2 \sqrt{K}\,, \end{aligned}$$

where the last inequality follows from the concavity of the square root. The final claim is obtained by bounding the diameter $D_{H_{1/2}}$ by $2\sqrt{K}$. □

We conclude this section by providing a proof of Lemma 4.3.

Proof of Lemma 4.3. The fact that $\delta_t^{\mathrm{Ts}} \geqslant 0$ holds actually for all regularizers and can be seen from the definition (4.25) with $p = p_t$. For the inequality $\delta_t^{\mathrm{Ts}} \leqslant M - m$, we start with elementary manipulations of the definition of the mixability gap (4.25). Denoting by $\vec{M}$ the vector with coordinates $(M, \dots, M)$ and noting that $\Big\langle p_t - q,\, \vec{M} \Big\rangle = 0$ for all $q \in \mathcal{S}$, we have

$$\delta_t^{\mathrm{Ts}} = \max_{q \in \mathcal{S}} \left\{ \langle p_t - q,\, -\widehat{y}_t \rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \right\} = \max_{q \in \mathcal{S}} \left\{ \Big\langle p_t - q,\, \vec{M} - \widehat{y}_t \Big\rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \right\}. \quad (4.42)$$

Since all the coordinates of $\vec{M} - \widehat{y}_t$ are non-negative and by non-negativity of the Bregman divergence, this implies that

$$\delta_t^{\mathrm{Ts}} \leqslant \Big\langle p_t,\, \vec{M} - \widehat{y}_t \Big\rangle = M - y_{A_t,t} \leqslant M - m\,.$$

We now prove the second inequality; we may assume that $\eta_t < +\infty$, as the bound holds trivially otherwise. By Proposition 4.2 (and by calculations similar to the ones performed in the proof of Reminder 4.5) the maximum in the rewriting (4.42) of δ_t^{Ts} is achieved on the interior of the domain of $H_{1/2}$, which equals $(0, +\infty)^K$, thus in the interior of $\mathcal{S}$. We therefore only need to prove that

$$\forall q \in \mathrm{Int}(\mathcal{S}), \qquad \Big\langle p_t - q,\, \vec{M} - \widehat{y}_t \Big\rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \leqslant \eta_t\, p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2\,. \quad (4.43)$$

We fix such a $q \in \mathrm{Int}(\mathcal{S})$, i.e., such that $q_a > 0$ for all a. We consider two cases. First, if $q_{A_t} \geqslant p_{t,A_t}$, then, given the observations made after (4.34),

$$\Big\langle p_t - q,\, \vec{M} - \widehat{y}_t \Big\rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} = \underbrace{\left(\frac{M - y_{t,A_t}}{p_{t,A_t}} \right)}_{\geqslant 0} \underbrace{\big(p_{t,A_t} - q_{A_t} \big)}_{\leqslant 0} - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \leqslant 0\,.$$

Otherwise, when $q_{A_t} < p_{t,A_t}$, a standard way of bounding the mixability gap, detailed below, indicates that

$$\langle p_t - q,\, M - \widehat{y}_t \rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \leqslant \frac{\eta_t}{2} \Big\langle \vec{M} - \widehat{y}_t,\, \nabla^2 H_{1/2}(z)^{-1} \big(\vec{M} - \widehat{y}_t \big) \Big\rangle\,, \quad (4.44)$$

where z is some probability distribution of the open segment $\mathrm{Seg}(q, p_t)$ between q and p_t, and where $\nabla^2 H_{1/2}(z)^{-1}$ denotes the inverse of the positive definite Hessian of $H_{1/2}$ at z. Since at $w \in (0, +\infty)^K$, the function $H_{1/2}$ is indeed twice differentiable, with

$$\nabla H_{1/2}(w) = \big(-w_a^{-1/2} \big)_{a \in [K]} \qquad \text{and} \qquad \nabla^2 H_{1/2}(w) = \mathrm{Diag}\big(w_a^{-3/2}/2 \big)_{a \in [K]}\,,$$

we have $\nabla^2 H_{1/2}(z)^{-1} = \mathrm{Diag}\big(2 z_a^{3/2} \big)_{a \in [K]}$. We substitute this value into (4.44) and recall that the vector $\vec{M} - \widehat{y}_t$ has null coordinates except for its A_t–th coordinate:

$$\frac{\eta_t}{2} \Big\langle \vec{M} - \widehat{y}_t,\, \nabla^2 H_{1/2}(z)^{-1} \big(\vec{M} - \widehat{y}_t \big) \Big\rangle = \eta_t\, z_{A_t}^{3/2} \big(M - \widehat{y}_{t,A_t} \big)^2\,.$$

Finally, remember that z lies in the open segment $\mathrm{Seg}(q, p_t)$ and that we assumed $q_{A_t} < p_{t,A_t}$; we thus also have $z_{A_t} < p_{t,A_t}$. As a consequence, using the very definition of $\widehat{y}_{t,A_t}$,

$$\eta_t\, z_{A_t}^{3/2} \big(M - \widehat{y}_{t,A_t} \big)^2 \leqslant \eta_t\, p_{t,A_t}^{3/2} \big(M - \widehat{y}_{t,A_t} \big)^2 = \eta_t\, p_{t,A_t}^{-1/2} (M - y_{t,A_t})^2\,.$$

Therefore, in all cases, that is, whether $q_{A_t} \geqslant p_{t,A_t}$ or $q_{A_t} < p_{t,A_t}$, the bound (4.43) is obtained. It only remains to prove the standard inequality (4.44).

This inequality is essentially stated as Therorem 26.13 in Lattimore and Szepesvári [2020] but we provide a proof for the sake of completeness. As we assumed that $\eta_t < +\infty$, we have (as above, by Proposition 4.2) that p_t lies in the interior of $\mathcal{S}$. In particular, as both p_t and q are in the interior of $\mathcal{S}$, the function $H_{1/2}$ is $\mathcal{C}^2$ over the closed segment $\overline{\mathrm{Seg}}(q, p_t)$ between q and p_t. Therefore, by the mean-value theorem, there exists z in the open segment $\mathrm{Seg}(q, p_t)$ such that

$$\underbrace{H_{1/2}(q) - H_{1/2}(p_t) - \big\langle \nabla H_{1/2}(p_t),\, q - p_t \big\rangle}_{= B_{H_{1/2}}(q, p_t)} = \frac{1}{2}\Big\langle q - p_t,\, \nabla^2 H_{1/2}(z)\,(q - p_t) \Big\rangle .$$

It is useful to introduce the standard notation from convex analysis for the local norm (which is indeed a norm because the Hessian is positive definite):

$$\|q - p_t\|^2_{\nabla^2 H_{1/2}(z)} \stackrel{\text{def}}{=} \Big\langle q - p_t,\, \nabla^2 H_{1/2}(z)\,(q - p_t) \Big\rangle .$$

We therefore have so far the rewriting:

$$-\frac{B_{H_{1/2}}(q, p_t)}{\eta_t} = -\frac{1}{2\eta_t}\Big\langle q - p_t,\, \nabla^2 H_{1/2}(z)\,(q - p_t) \Big\rangle .$$

Now, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\Big\langle p_t - q,\, \vec{M} - \widehat{y}_t \Big\rangle &= \Big\langle \nabla^2 H_{1/2}(z)^{1/2}\,(p_t - q),\, \nabla^2 H_{1/2}(z)^{-1/2}\,\big(\vec{M} - \widehat{y}_t\big) \Big\rangle \\ &\leqslant \|p_t - q\|_{\nabla^2 H_{1/2}(z)}\ \|\vec{M} - \widehat{y}_t\|_{\nabla^2 H_{1/2}(z)^{-1}} .\end{aligned}$$

Combining the rewriting and the bound above, we get

$$\begin{aligned}&\langle p_t - q,\, M - \widehat{y}_t \rangle - \frac{B_{H_{1/2}}(q, p_t)}{\eta_t} \\ &\leqslant \|p_t - q\|_{\nabla^2 H_{1/2}(z)} \|\vec{M} - \widehat{y}_t\|_{\nabla^2 H_{1/2}(z)^{-1}} - \frac{1}{2\eta_t}\|q - p_t\|^2_{\nabla^2 H_{1/2}(z)} \\ &\leqslant \frac{\eta_t}{2}\|\vec{M} - \widehat{y}_t\|^2_{\nabla^2 H_{1/2}(z)^{-1}} ,\end{aligned}$$

where we used $ab - b^2/2 \leqslant a^2/2$ to get the second inequality. This is exactly (4.44). □

4.D. Adaptation to the range for linear bandits

To illustrate the generality of the techniques discussed in this chapter, we quickly describe how these can be used to obtain range adaptive algorithms for linear bandits. This section is meant for illustration and not for completeness. In particular, we focus on the case of (oblivious) adversarial linear bandits, for which we refer the reader to Lattimore and Szepesvári [2020, Chapter 27], which we follow closely, for a more thorough description of the setting; we do not describe the application of our techniques to stochastic linear bandits.

Learning protocol. A finite action set $\mathcal{A} \subset \mathbb{R}^d$, of cardinality K, is given. (The setting of vanilla K–armed bandits considered in the rest of the chapter corresponds to $\mathcal{A}$ formed by the vertices of the probability simplex of $\mathbb{R}^K$.) The environment selects beforehand a sequence $(y_t)_{t\geqslant 1}$ of vectors in $\mathbb{R}^d$ satisfying a boundedness assumption: there exists an interval $[m, M]$ such that

$$\forall t \geqslant 1, \quad \forall x \in \mathcal{A}, \qquad x^\top y_t \in [m, M]\,. \tag{4.45}$$

We assume that the player does not know in advance m nor M. To simplify the exposition, we also assume that $m \leqslant 0 \leqslant M$.

At every time step, the player chooses an action $X_t \in \mathcal{A}$ and receives and only observes the payoff $X_t^\top y_t$. It does not observe y_t nor the payoffs $x^\top y_t$ associated with choices $x \neq X_t$. The action X_t is chosen independently at random according to a distribution over $\mathcal{A}$ denoted by $p_t = \big(p_t(a)\big)_{a\in\mathcal{A}}$.

The expected regret is defined as

$$R_T(y_{1:T}) = \max_{x\in\mathcal{A}} \sum_{t=1}^T x^\top y_t - \mathbb{E}\left[\sum_{t=1}^T X_t^\top y_t\right].$$

Estimating the unobserved payoffs. As in the case of vanilla K–armed bandits, the key is to estimate unobserved payoffs. We may actually build an estimate $\widehat{y}_t$ of the vectors y_t, from which we form the estimates $x^\top \widehat{y}_t$. This estimate takes advantage of the linear structure of the problem.

Fix a distribution π such that the non-negative symmetric matrix

$$M(\pi) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{A}} \pi(x)\, xx^\top$$

is invertible: such a distribution exists whenever $\mathcal{A}$ spans $\mathbb{R}^d$, which we may assume with no loss of generality; see Lemma 4.4 below. This distribution π will be used to explore the arms; it is in general not uniform over the arms. For all distributions q over $\mathcal{A}$ and all $\gamma \in (0, 1]$, the distribution $p = (1-\gamma)q + \gamma\pi$ is such that the non-negative symmetric matrix $M(p)$ is invertible as well (as it is larger than $\gamma\, M(\pi)$, in the sense of the partial inequality $\succcurlyeq$ over non-negative symmetric matrices). We then define

$$\widehat{y}_t = M(p_t)^{-1}\, X_t X_t^\top y_t \tag{4.46}$$

and note that

$$\mathbb{E}\big[\widehat{y}_t \,\big|\, p_t\big] = M(p_t)^{-1} \Bigg(\underbrace{\sum_{x\in\mathcal{A}} p_t(x)\, xx^\top}_{=M(p_t)}\, y_t \Bigg) = y_t\,; \tag{4.47}$$

indeed, conditioning to p_t amounts to integrating over the random choice of X_t according to p_t.

Algorithm 4.4 AdaHedge for adversarial linear bandits

Input: an exploration distribution π over $\mathcal{A}$ and exploration rates $(\gamma_t)_{t\geqslant 1}$ in $[0,1]$
Initialization: $\eta_1 = +\infty$ and q_1 is the uniform distribution over $\mathcal{A}$
for rounds $t = 1, \dots$ **do**
  Define p_t by mixing q_t with π according to

$$p_t = (1-\gamma_t)q_t + \gamma_t \pi$$

  Draw an arm $X_t \sim p_t$ (independently at random according to the distribution p_t)
  Get and observe the payoff $X_t^\top y_t$
  Compute estimates $x^\top \widehat{y}_t$ of all payoffs according to (4.46)
  Compute the mixability gap δ_t based on the distribution q_t and on these estimates:

$$\delta_t = \begin{cases} -\sum_{x\in\mathcal{A}} q_t(x)\, x^\top \widehat{y}_t + \dfrac{1}{\eta_t} \ln\left(\sum_{x\in\mathcal{A}} q_t(x) \mathrm{e}^{\eta_t x^\top \widehat{y}_t}\right) & \text{if } \eta_t < +\infty \\ -\sum_{x\in\mathcal{A}} q_t(x)\, x^\top \widehat{y}_t + \max_{x\in\mathcal{A}} x^\top \widehat{y}_t & \text{if } \eta_t = +\infty \end{cases}$$

  Compute the learning rate $\eta_{t+1} = \left(\sum_{s=1}^{t} \delta_s\right)^{-1} \ln K$
  Define q_{t+1} component-wise as

$$q_{t+1}(a) = \exp\left(\eta_{t+1} \sum_{s=1}^{t} a^\top \widehat{y}_s\right) \Big/ \sum_{x\in\mathcal{A}} \exp\left(\eta_{t+1} \sum_{s=1}^{t} x^\top \widehat{y}_s\right)$$

end for

An algorithm adaptive to the unknown range. When the range is given, a well-known strategy is to use plain exponential weights over actions in $\mathcal{A}$ with the estimates $x^\top \widehat{y}_t$ to obtain distributions q_t that are then mixed with π to form the final distributions p_t. When the range is unknown, we suggest to simply replace plain exponential weights with AdaHedge (the difference lies in the tuning of the rates η_t), which leads to Algorithm 4.4. In this algorithm, we refer to rates γ_t as exploration rates (and not as extra-exploration rates as in Algorithm 4.1) and similarly, to π as the exploration distribution. This is because for adversarial linear bandits, exploration was always required even to get expected results (unlike for K–armed bandits, see the introduction of Section 4.5).

The analysis of this algorithm relies on the same ingredients as the ones already encountered in Section 4.5.1, with the addition of the following lemma, that quantifies the quality of the exploration. This lemma requires that $\mathcal{A}$ spans $\mathbb{R}^d$, which we may assume with no loss of generality (otherwise, we just replace $\mathbb{R}^d$ by the vector space generated by $\mathcal{A}$).

Lemma 4.4 (Lattimore and Szepesvári [2020, Theorem 21.1]). *There exists a distribution π over $\mathcal{A}$ such that*

$$M(\pi) = \sum_{x\in\mathcal{A}} \pi(x)\, xx^\top \text{ is invertible} \qquad \text{and} \qquad \max_{x\in\mathcal{A}} x^\top M(\pi)^{-1} x = d\,.$$

We are now ready to state the main result of this section. It is the counterpart of Corollary 4.1; for the sake of simplicity, we only state it for the value $\alpha = 1/2$.

Theorem 4.9. *AdaHedge for adversarial linear bandits (Algorithm 4.4) with the extra-exploration*

$$\gamma_t = \min\Big\{1/2,\, \sqrt{2.5\, d(\ln K) t^{-1/2}}\Big\}$$

ensures that for all bounded ranges $[m, M]$, for all oblivious individual sequences $y_1, y_2, \dots$ satisfying the boundedness condition (4.45)*,*

$$R_T(y_{1:T}) \leqslant 12(M-m)\sqrt{dT \ln K} + 18(M-m)d \ln K\,.$$

The proof starts by following closely the ones of Theorem 4.3 and Corollary 4.1; the differences are underlined and dealt with in the second part of the proof.

Proof. By Reminder 4.2, since the player plays the AdaHedge strategy over the payoff estimates $x^\top \widehat{y}_t$, the pre-regret satisfies

$$\max_{x\in\mathcal{A}} \sum_{t=1}^T x^\top \widehat{y}_t - \sum_{t=1}^T \sum_{a\in\mathcal{A}} q_t(a)\, a^\top \widehat{y}_t \leqslant 2\sqrt{V_T \ln K} + M_T \left(2 + \frac{4}{3}\ln K\right)$$

with $V_T = \displaystyle\sum_{t=1}^T \sum_{x\in\mathcal{A}} q_t(x) \left(x^\top \widehat{y}_t\right)^2$ and

$$M_T = \max\big\{x^\top \widehat{y}_t : t \leqslant T \text{ and } x \in \mathcal{A}\big\} - \min\big\{x^\top \widehat{y}_t : t \leqslant T \text{ and } x \in \mathcal{A}\big\}\,.$$

As in Theorem 4.3, since $\gamma_t \leqslant 1/2$, we have $q_t(x) \leqslant 2\, p_t(x)$ for all $x \in \mathcal{A}$. We therefore define

$$V'_T = \sum_{t=1}^T \sum_{x\in\mathcal{A}} p_t(x) \left(x^\top \widehat{y}_t\right)^2$$

and have $V_t \leqslant 2V'_T$. By the tower rule, based on the equality (4.47), and given that the expectation of a maximum is larger than the maximum of the expectations (for the first inequality), and by the definition of the p_t (for the second inequality), we have proved so far that

$$\begin{aligned}
R_T(y_{1:T}) &\leqslant \mathbb{E}\left[\max_{x\in\mathcal{A}} \sum_{t=1}^T x^\top \widehat{y}_t - \sum_{t=1}^T \sum_{a\in\mathcal{A}} p_t(a)\, a^\top \widehat{y}_t\right] \\
&\leqslant \mathbb{E}\left[\max_{x\in\mathcal{A}} \sum_{t=1}^T x^\top \widehat{y}_t - \sum_{t=1}^T \sum_{a\in\mathcal{A}} q_t(a)\, a^\top \widehat{y}_t\right] + \mathbb{E}\left[\sum_{t=1}^T \gamma_t \sum_{a\in\mathcal{A}} \big(\pi(a) - q_t(a)\big)\, a^\top \widehat{y}_t\right] \\
&\leqslant \mathbb{E}\left[2\sqrt{2V'_T \ln K} + M_T\left(2 + \frac{4}{3}\ln K\right)\right] + \sum_{t=1}^T \gamma_t \underbrace{\sum_{a\in\mathcal{A}} \big(\pi(a) - q_t(a)\big)\, a^\top y_t}_{\leqslant (M-m)}\,.
\end{aligned}$$

Hence by Jensen's inequality and by the bounds $\mathbb{E}[V'_T] \leqslant (M-m)^2 dT$ and $M_T \leqslant 2(M-m)d/\gamma_T$ proved below, we finally get

$$\begin{aligned}
R_T(y_{1:t}) &\leqslant 2\sqrt{2\,\mathbb{E}[V'_T] \ln K} + \mathbb{E}[M_T]\left(2 + \frac{4}{3}\ln K\right) + (M-m)\sum_{t=1}^T \gamma_t \\
&\leqslant 2\sqrt{2}(M-m)\sqrt{dT \ln K} + \left(2 + \frac{4}{3}\ln K\right)\frac{2(M-m)d}{\gamma_T} + (M-m)\sum_{t=1}^T \gamma_t \\
&\leqslant 3(M-m)\sqrt{dT \ln K} + 9(M-m)\frac{d \ln K}{\gamma_T} + (M-m)\sum_{t=1}^T \gamma_t\,.
\end{aligned}$$

Replacing the γ_t by their values and using the same bounds as in Corollary 4.1 yields the claimed result; the factor 12 in the bound comes from

$$3 + \sqrt{10} + 9\sqrt{\frac{2}{5}} \leqslant 12\,.$$

We only need to prove the two claimed bounds to complete the proof; they can be extracted from the proof of Theorem 27.1 by Lattimore and Szepesvári [2020] but we provide derivations for the sake of completeness.

Proof of $M_T \leqslant 2(M-m)d/\gamma_T$. We fix $x \in \mathcal{A}$ and $t \leqslant T$. We recall that $M(p_t)$ and thus $M(p_t)^{-1}$ are positive definite symmetric matrices. By the Cauchy-Schwarz inequality applied with the norm induced by the positive $M(p_t)^{-1}$,

$$\left|x^\top M(p_t)^{-1} X_t\right| \leqslant \sqrt{x^\top M(p_t)^{-1} x}\,\sqrt{X_t^\top M(p_t)^{-1} X_t} \leqslant \max_{x\in\mathcal{A}} \left\{x^\top M(p_t)^{-1} x\right\}.$$

As indicated right before (4.46), we have $M(p_t) \succcurlyeq \gamma_t\, M(\pi)$ and therefore $M(p_t)^{-1} \preccurlyeq M(\pi)^{-1}/\gamma_t$. This entails

$$\left|x^\top M(p_t)^{-1} X_t\right| \leqslant \frac{1}{\gamma_t} \max_{x\in\mathcal{A}} \left\{x^\top M(\pi)^{-1} x\right\} = \frac{d}{\gamma_t} \leqslant \frac{d}{\gamma_T}\,,$$

where the equality follows from Lemma 4.4 and where we used $\gamma_T \leqslant \gamma_t$ for the second inequality. Finally, keeping in mind that we assumed $m \leqslant 0 \leqslant M$,

$$x^\top \widehat{y}_t = \underbrace{x^\top M(p_t)^{-1}\, X_t}_{\in[-d/\gamma_t, d/\gamma_t]}\, \underbrace{X_t^\top y_t}_{\in[m,M]} \in \left[-\frac{d\max\{-m,M\}}{\gamma_T}, \frac{d\max\{-m,M\}}{\gamma_T}\right],$$

from which the bound

$$M_t = 2\,\frac{d\max\{-m,M\}}{\gamma_T} \leqslant \frac{2d(M-m)}{\gamma_T}$$

follows, as desired.

Proof of $\mathbb{E}[V_T'] \leqslant (M-m)^2 dT$. Since $\left|X_t^\top y_t\right| \leqslant \max\{-m,M\} \leqslant M-m$, the definition (4.46) leads to

$$\begin{aligned}\left(x^\top \widehat{y}_t\right)^2 = \left(x^\top M(p_t)^{-1}\, X_t X_t^\top y_t\right)^2 &\leqslant (M-m)^2 \left(x^\top M(p_t)^{-1}\, X_t\right)^2 \\ &= (M-m)^2\, X_t^\top M(p_t)^{-1} x x^\top M(p_t)^{-1} X_t\,.\end{aligned}$$

Therefore, summing over $x \in \mathcal{A}$ and using the very definition of $M(p_t)$, we get

$$\begin{aligned}\sum_{x\in\mathcal{A}} p_t(x)\left(x^\top \widehat{y}_t\right)^2 &\leqslant (M-m)^2\, X_t^\top M(p_t)^{-1} \left(\sum_{x\in\mathcal{A}} p_t(x)\, x x^\top\right) M(p_t)^{-1} X_t \\ &= (M-m)^2\, X_t^\top M(p_t)^{-1} X_t = (M-m)^2 \operatorname{Tr}\Big(M(p_t)^{-1} X_t X_t^\top\Big)\,.\end{aligned}$$

Now, by the linearity of the trace,

$$\mathbb{E}\left[\operatorname{Tr}\Big(M(p_t)^{-1} X_t X_t^\top\Big)\right] = \mathbb{E}\left[\sum_{x\in\mathcal{A}} p_t(x) \operatorname{Tr}\Big(M(p_t)^{-1} x x^\top\Big)\right] = \mathbb{E}\big[\operatorname{Tr}(I_d)\big] = d\,,$$

where I_d is the d–dimensional identity matrix. Collecting all bounds together and summing over t yields the claimed inequality $\mathbb{E}[V_T'] \leqslant (M-m)^2 dT$. □

Chapter 5.

Diversity-preserving bandits, revisited

Abstract

We consider the bandit-based framework for diversity-preserving recommendations discussed in Celis et al. [2019]. We design algorithms using the specific structure of the setting; ; they are variants of UCB and of follow-the-leader approaches. These algorithms are efficient and enjoy low regret, while naturally satisfying the diversity-preserving constraints. We carry out a detailed analysis, providing minimax and distribution-dependent regret bounds; we also uncover the possibility of bounded regret for some specific action sets. This analysis is also supported with lower bounds on the regret, proving that our results are unimprovable in general. Experiments on synthetic data illustrate the performance of our algorithms.

This chapter is based on ongoing work, in collaboration with Sébastien Gerchinovitz, Jean-Michel Loubes and Gilles Stoltz. The recent preprint Hadiji et al. [2020] (under review) was built upon this chapter.

Contents

5.1. Setting and literature review

We consider stochastic bandit models with finitely many arms. All of them are desirable actions, though some lead to higher payoffs. Effective (regret-minimizing) algorithms are bound to play the optimal arm(s) an overwhelming fraction of time. Celis et al. [2019] refer to this effect as polarization and introduce a model to avoid it. We suggest the alternative terminology of preserving diversity. A general formulation of the bandit model by Celis et al. [2019] is provided below and is summarized in Protocol 5.1. Our aim in this chapter is to deepen and improve on the results obtained by the mentioned reference; see Section 5.1.3 for details.

Diversity-preserving bandits, as introduced by Celis et al. [2019]. As in traditional K–armed bandits, probability distributions $\nu_1, \dots, \nu_K$ associated with each arm are considered, with expectations denoted by $\mu_1, \dots, \mu_K$. These distributions are unknown to the learner but belong to a known set of possible distributions, called a model $\mathcal{D}$.

The learning protocol is the following. An arm $A_t \in [K]$ is picked among K choices at each round, where we denote by $[K]$ the set $\{1, \dots, K\}$. The learner then obtains a payoff Y_t drawn independently at random according to ν_{A_t} given that choice. This is the only observation made (the learner does not know what it would have obtained with a different choice). However, the distinguishing feature of the bandit model by Celis et al. [2019] is that the choice of A_t is made in two steps, as follows. Denote by $\mathcal{S}$ the set of distributions over the K arms. First, a distribution $\underline{p}_t \in \mathcal{S}$ is picked, in some known closed set $\mathcal{P}$, which quantifies diversity (specific examples are given below). Then, the arm A_t is drawn independently at random according to $\underline{p}_t$. Following game-theoretic terminology, we will call $a \in [K]$ pure actions or ams, and $\underline{p} \in \mathcal{P}$ mixed actions or probabilities.

Protocol 5.1 Diversity-preserving stochastic bandits (Celis et al., 2019)

Known parameters
Number K of arms
Model $\mathcal{D}$ of possible distributions
Closed set $\mathcal{P}$ of diverse enough probability distributions over the arms
Unknown parameters
Probability distributions $\nu_1, \dots, \nu_K \in \mathcal{D}$ for each arm, with expectations $\underline{\mu} = (\mu_1, \dots, \mu_K)$

for $t = 1, 2, \dots$ **do**
Pick a distribution $\underline{p}_t = (p_{t,1}, \dots, p_{t,K}) \in \mathcal{P}$ over the arms
Draw independently at random an arm $A_t \sim \underline{p}_t$
Get and observe a payoff $Y_t \sim \nu_{A_t}$ drawn independently at random according to ν_{A_t} given A_t
end for

Aim
Minimize the expected regret $\quad R_T = T \max_{\underline{p} \in \mathcal{P}} \langle \underline{p}, \underline{\mu} \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle \underline{p}_t, \underline{\mu} \rangle \right]$

We measure performance in terms of expected payoffs. The expected payoff at round t may be computed by repeated applications of the tower rule:

$$\mathbb{E}[Y_t \mid A_t] = \mu_{A_t}\,, \quad \text{thus} \quad \mathbb{E}\big[Y_t \mid \underline{p}_t\big] = \sum_{k \in [K]} p_{t,k}\, \mu_k \overset{\text{def}}{=} \langle \underline{p}_t, \underline{\mu} \rangle\,, \quad \text{thus} \quad \mathbb{E}[Y_t] = \mathbb{E}\big[\langle \underline{p}_t, \underline{\mu} \rangle\big]\,.$$

We denoted by $\langle \underline{p}_t\,, \underline{\mu} \rangle$ the inner product between the vectors $\underline{p}_t$ and $\underline{\mu}$. Maximizing the cumulative

expected payoff of a policy amounts to minimizing the expected regret defined as

$$R_T = T \max_{\underline{p} \in \mathcal{P}} \langle \underline{p}, \underline{\mu} \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle \underline{p}_t, \underline{\mu} \rangle \right].$$

In the definition of the regret, the comparison is made with respect to the expected payoff that would have been obtained by picking at each round a best diversity-preserving distribution over the arms.

Bandit model. In this chapter we consider mainly the bandit model $\mathcal{D}_{[0,1]}$ of probability measures supported on $[0,1]$, that is, we assume that rewards can be distributed according to any distribution bounded in $[0,1]$. An exception to this is the lower bound in Section 5.5, which we formulate on a generic model $\mathcal{D}$.

5.1.1. Examples of diversity-preserving sets $\mathcal{P}$ of distributions over the arms

Simplest example. The simplest requirement is that each arm should be played with some minimal probability $\ell > 0$, which corresponds to

$$\mathcal{P} = \left\{ \underline{p} : \quad \forall a \in [K], \quad p_a \geqslant \ell \right\}.$$

More generally, Celis et al. [2019] indicate that one could group arms intro groups $G_1, \dots, G_N$ of similar arms and impose minimal probabilities $\ell_1, \dots, \ell_N > 0$ as well as maximal probabilities $u_1, \dots, u_N < 1$ for each group defined as:

$$\mathcal{P} = \left\{ \underline{p} : \quad \forall g \in [N], \quad \sum_{a \in G_g} p_a \in [\ell_g, u_g] \right\}.$$

The sets $\mathcal{P}$ considered above are polytopes.

General polytopes. To justify more general types of probability sets $\mathcal{P}$, we could envision a setting in which each every pure action a has N costs $c_a^{(1)}, \dots, c_a^{(N)}$ in $\mathbb{R}$. These costs could represent for example some limited resources, or some environmental cost like the amount of carbon emissions generated from taking the action. The model can handle negative costs, e.g., negative carbon emissions. The name "diversity-preserving" was inspired by the example of the previous paragraph, and is perhaps less pertinent in the present example.

Then, if the player picks an action at random according to some probability $(p_1, \dots, p_K)$ over the set of actions $[K]$, the N expected costs of her choice are

$$\sum_{a=1}^{K} p_a c_a^{(1)} \ , \ \dots \ , \ \sum_{a=1}^{K} p_a c_a^{(N)} .$$

In this case a reasonable objective for the player is to maximize her payoff under the constraints that, for all $i \in [N]$, the i–th expected cost of her actions be kept under a certain level u_i, or above a certain level ℓ_i in case of negative costs. This amounts to playing under Protocol 5.1, with the probability set

$$\mathcal{P} = \left\{ \underline{p} : \quad \forall n \in [N], \quad \sum_{a=1}^{K} p_a c_a^{(n)} \in [\ell_n, u_n] \right\}.$$

These sets are again polytopes, and generalize the previous example.

Other examples, and the number of vertices of $\mathcal{P}$. We argue in Open question 5.1 that some naive approaches could yield regret bounds that scale linearly with the number of vertices of the polytope $\mathcal{P}$, making the number of vertices an important property of $\mathcal{P}$. In general, it is difficult to compute the number of vertices of a polytope, and there can be arbitrarily many vertices. However, in some simple cases, the number is manageable, perhaps making the naive approach relevant again. We discuss a few examples here.

For instance, consider the simplest diversity-preserving example, with ℓ a diversity-preserving threshold,

$$\mathcal{P} = \left\{\underline{p} : \quad \forall a \in [K], \quad p_a \geqslant \ell\right\}.$$

This set has K vertices if $\ell \in [0, 1/K)$ and is empty if $\ell > 1/K$ (and one vertex when $\ell = 1/K$). Another related example is the following probability set; fix $u > 0$ a threshold and define

$$\mathcal{P} = \left\{\underline{p} : \quad \forall a \in [K], \quad p_a \leqslant u\right\}.$$

In this case u acts as a parameter preventing polarization. This probability set can have up to $K(K-1)$ vertices, if $u \in (1/2, 1]$. Indeed, when $u \in (1/2, 1]$, the vertices are of the form $(0, \dots, u, \dots, 1-u, \dots, 0)$, with u and $1-u$ at any pair of coordinates and 0 elsewhere; there are $K(K-1)$ such vertices.

In these two cases, the number of vertices grows, respectively, linearly and quadratically with the ambient dimension K. Since we typically assume that K is of manageable size, this means that regret bounds scaling linearly with the number of vertices are not prohibitive.

Let us consider another example. Denote by $S_{K'}$ the K' dimensional simplex. Choose $\mathcal{P}_1 \subset \mathcal{S}_{K_1}$ and $\mathcal{P}_2 \subset \mathcal{S}_{K_2}$, and some weights $\pi = (\pi_1, \pi_2)$ such that $\pi_1 > 0$ and $\pi_2 > 0$ and $\pi_1 + \pi_2 = 1$ (the case when $\pi_1 = 0$ or $\pi_2 = 0$ is not of interest). Then define the π-weighed product of $\mathcal{P}_1$ and $\mathcal{P}_2$ to be

$$\pi(\mathcal{P}_1, \mathcal{P}_2) = \left\{(\underline{p}^{(1)}, \underline{p}^{(2)}) \in \mathcal{S}_{K_1+K_2} \ : \ \forall j \in \{1,2\}, \quad \sum_{i=1}^{K_j} p_i^{(j)} = \pi_j \quad \text{and} \quad \frac{1}{\pi_j}\underline{p}^{(j)} \in \mathcal{P}_j\right\}$$

Then the number of vertices of $\pi(\mathcal{P}_1, \mathcal{P}_2)$ is the product of the number of vertices of $\mathcal{P}_1$ and $\mathcal{P}_2$. Generalizing this construction to a larger number of sets, we obtain a natural yet non-trivial example of a probability set with a number of vertices growing exponentially with the dimension.

Literature review on fairness and diversity in stochastic bandits. In the line of work by Joseph et al. [2016], Amani et al. [2019], Liu et al. [2017] and Gillen et al. [2018], the learner wishes that its actions satisfy certain constraints with high probability. These constraints are inspired by the framework of individual fairness that states that similar individuals should be treated similarly—there, actions correspond to individuals. In these models, the constraints depend on the unknown problem, and therefore the usual tradeoff between exploration and exploitation is modified: the player needs to explore some more in order to learn the constraint while playing the bandit game. This is mathematically quite different from our setting.

As far as diversity is concerned, Li et al. [2019] consider problem called combinatorial sleeping bandits, in which the player may pick multiple actions among the K available at every step. The authors indeed impose that their algorithms satisfy a diversity preserving condition on the choice of the actions, but this condition is only asymptotical. Patil et al. [2019] propose another bandit framework in the same vein. They derive bandit algorithms that ensure that the proportion of times each action is selected is lower bounded, i.e., with our notation that $N_a(T)/T \geqslant \alpha$ almost surely. Although the objective is similar in spirit, this constraint leads to design issues for the algorithm that are quite different from ours, and are arguably less mathematically elegant. For instance, in their setting, pulling the first arm automatically violates the fairness constraints as

the proportion $N_a(T)/T$ for the arms not pulled is 0, thus the constraint can only be enforced for T large enough. This also leads to crucial differences between horizon dependent and anytime algorithms. Our setting enforces similar guarantees while bypassing these isssues.

5.1.2. Comparison to stochastic linear bandits

As noted by Celis et al. [2019], the setting considered is a special case of linear stochastic bandits (see below for an extensive literature review on this matter). Indeed, the expected payoff obtained at each round equals $\mathbb{E}\big[\langle \underline{p}_t, \underline{\mu} \rangle\big]$, which is a linear function of the probability $\underline{p}_t$ picked. This observation opens the toolbox of algorithms to deal with stochastic linear bandits (with action set $\mathcal{A} = \mathcal{P}$) to solve the considered problem; this is exactly the approach followed by Celis et al. [2019].

However, doing so, one discards the pure action A_t picked, which is known, and one relates the reward $Y_t \sim \mu_{A_t}$ to $\underline{p}_t$ and not to A_t, which is a loss of information. We show in this chapter that by taking the intermediate pure action A_t into account, instead of only considering the mixed action $\underline{p}_t$, that sharper regret bounds than the ones of Celis et al. [2019] may be achieved; see below the intuition behind this fact.

The considered setting can thus be described as a stochastic linear bandit setting with augmented feedback. Other bandit models with additional feedback (and thus, improved bounds) have been studied, e.g., in Caron et al. [2012] and Degenne et al. [2018]. More recently, Kirschner et al. [2020] considered the general linear partial monitoring problem, in which the learner observes a linear functional depending on the chosen action of the parameter $\vec{\mu}$, perturbed by some noise, instead of the reward. While the problematic is similar to ours, the settings are formally independent.

Intuition behind the possibility of sharper bounds. Let us measure, through Kullback-Leibler divergences (denoted by KL), the information available when discriminating between two bandit problems $\underline{\nu} = (\nu_1, \dots, \nu_K)$ and $\underline{\nu}' = (\nu'_1, \dots, \nu'_K)$, depending on whether the specific arm A_t is observed or not. Under the problem $\underline{\nu}$ and conditionally to the choice of a distribution $\underline{p}_t$ over the arms, the learner sees the payoff Y_t as distributed according to some unconditional distribution when A_t is not taken into account, and the conditional distribution ν_{A_t} when A_t is taken into account:

$$Y_t \sim \sum_{a \in [K]} p_{t,a} \nu_a \qquad \text{and} \qquad Y_t \,|\, A_t \sim \nu_{A_t} \,,$$

respectively. Conditionally to the choice of $\underline{p}_t$, the Kullback-Leibler divergences between the distributions of Y_t under $\underline{\nu}$ and $\underline{\nu}'$ are therefore given by

$$\underbrace{\mathrm{KL}\left(\sum_{a \in [K]} p_{t,a} \nu_a, \sum_{a \in [K]} p_{t,a} \nu'_a \right)}_{\text{without } A_t} \leqslant \underbrace{\mathbb{E}\big[\mathrm{KL}(\nu_{A_t}, \nu'_{A_t})\big]}_{\text{with } A_t} = \sum_{a \in [K]} p_{t,a} \,\mathrm{KL}(\nu_a, \nu'_a) \,,$$

where the inequality is by convexity of KL.

As we shall see Section 5.5, this technical observation becomes central when one wishes to derive lower bounds on the regret.

Regret bounds typically achieved by linear bandit algorithms. Let us recall here the linear bandit setting (see the monograph by Lattimore and Szepesvári, 2020, Chapter 19, for a longer description). An action set $\mathcal{A} \subset \mathbb{R}^d$ is given to the learner. Some parameter $\vec{\mu} \in \mathbb{R}^d$ is set but

remains unknown to the learner. The latter selects at each step an action $X_t \in \mathcal{A}$ and gets and observes a random reward Y_t such that $\mathbb{E}[Y_t \,|\, X_t] = \langle X_t,\, \vec{\mu} \rangle$. The expected regret is defined as

$$R_T^{\text{lin}} = T \max_{x \in \mathcal{A}} \langle x,\, \vec{\mu} \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t,\, \vec{\mu} \rangle \right].$$

We indicate below some typical regret bounds achieved in linear bandits, which will be used as benchmarks to compare our new bounds to.

A first stream of the literature focuses on generalizations of the UCB algorithm called LinUCB (linear upper confidence bound) or OFUL (optimism in the face of uncertainty for linear bandits); they were introduced by Li et al. [2010] and Chu et al. [2011] and studied by Abbasi-Yadkori et al. [2011]. They consider the set $\mathcal{L}$ of bandit models such that the parameter $\vec{\mu}$ satisfies $\langle x,\, \vec{\mu} \rangle \in [-1, 1]$ for all $x \in \mathcal{A}$ and the noise $Y_t - \mathbb{E}[Y_t | X_t]$ is sub-Gaussian (with constant less than $1/4$, say). The first kind of results they obtain is a distribution-free bound: for some numerical constant c,

$$\sup_{\mathcal{L}} R_T^{\text{lin}} \leqslant c\, d \sqrt{T} \ln T\,.$$

They also obtain finite-time distribution-dependent bounds in the case $\mathcal{A}$ is finite or is a polytope; we denote by $\mathcal{A}_{\text{finite}}$ the set of extremal points of $\mathcal{A}$ when $\mathcal{A}$ is a polytope (they generate $\mathcal{A}$) and $\mathcal{A}_{\text{finite}} = \mathcal{A}$ when $\mathcal{A}$ is finite. These finite-time distribution-dependent bounds are of the form: there exists a numerical constant C such that for each bandit model in $\mathcal{L}$,

$$R_T^{\text{lin}} \leqslant C\, \frac{1}{\Delta} \big(\ln^2 T + d \ln T + d^2 \ln \ln T \big)\,,$$

where the gap $\Delta(x)$ of an action $x \in \mathcal{A}$ and the overall gap Δ among suboptimal actions are defined as

$$\Delta(x) = \max_{y \in \mathcal{A}} \langle y - x,\, \vec{\mu} \rangle \qquad \text{and} \qquad \Delta = \min\big\{ \Delta(x) :\ x \in \mathcal{A}_{\text{finite}} \text{ s.t. } \Delta(x) > 0 \big\}\,.$$

A second stream of the literature improves on the treatment of the situation where $\mathcal{A}$ is finite or is a polytope and obtains distribution-dependent bounds that only scale with $\ln T$. Actually, such bounds could have been obtained by playing a plain UCB on $\mathcal{A}_{\text{finite}}$, but they would not get the optimal constant in front of the $\ln T$ (for more details about this suboptimality phenomenon, see Lattimore and Szepesvári, 2017). That is, asymptotically optimal distribution-dependent bounds are achieved: there exist algorithms such that for all bandit models in $\mathcal{L}$,

$$\limsup_{T \to \infty} \frac{R_T^{\text{lin}}}{\ln T} \leqslant \kappa(\mathcal{A}, \vec{\mu})\,,$$

where no reasonable algorithm can improve on the constant $\kappa(\mathcal{A}, \vec{\mu})$ when the noise is Gaussian; see Lattimore [2017], Combes et al. [2017], and Hao et al. [2020]. Further details on these results will be provided in Sections 5.2.5 and 5.5.

5.1.3. Summary of our contributions and outline of the chapter

In this chapter, our objective is to study how difficult it is to control the regret under the diversity-preserving constraint, i.e., how optimal regret bounds vary depending on P. Towards this goal, an essential element to understand is the influence of the geometry of the diversity preserving set $\mathcal{P}$. Our contributions in this respect can be separated into two parts: Sections 5.2 and 5.3 are devoted to upper bounds on the regret, with two new algorithms, while Section 5.4 and 5.5 provide lower bounds.

The first algorithm, in Section 5.2 is based on the standard UCB strategy, while the second, in Section 5.3, uses the follow-the-regularized-leader framework. Our two new algorithms for the diversity-preserving setting take full advantage of the specific observation protocole : they both behave provably better than linear bandit algorithms that do not exploit the diversity-preserving structure.

We analyze in depth the first algorithm inspired from UCB, providing various regret bounds: a general distribution-free regret bound of order $\mathcal{O}\big(\sqrt{KT\ln T}\big)$ in Section 5.2.2 and, in Section 5.2.3, a distribution-dependent regret bound when $\mathcal{P}$ is a polytope, of order $\mathcal{O}\big(\ln^2(T)/\Delta\big)$, where Δ is the minimal suboptimality gap among vertices of $\mathcal{P}$. Avenues for improvements are discussed in Section 5.2.4. We also prove a striking property of our algorithm in Section 5.2.5: it enjoys bounded regret when the diversity-preserving set $\mathcal{P}$ is included in the interior of the simplex.

Section 5.3 is devoted to our second algorithm. We show that it also achieves close to optimal distribution-free regret, and this, even in an adversarial setting (Section 5.3.4). We also provide an improved high-probability regret bound when $\mathcal{P}$ is included in the interior of the simplex (Section 5.3.3).

We also discuss the optimality of our approaches (and lack thereof) by providing two lower bounds on the regret suffered by any algorithm. Section 5.4 contains a distribution-free lower bound, which can be as large as $\Omega\big(\sqrt{KT}\big)$ for some probability sets, proving the quasi-optimality of our algorithms for these action sets. In Section 5.5, we describe in depth an asymptotic distribution-dependent lower bound on the regret, when $\mathcal{P}$ is a polytope. This lower bound, which is logarithmic in T, is expressed via an optimization problem.

Finally, we conclude this chapter with some numerical experiments (Section 5.6).

5.2. A UCB-like algorithm and its analysis

In this section we propose a simple variation of the well-known UCB algorithm designed for our problem in the bounded rewards model. After a presentation of the algorithm, we give three different regret bounds for this algorithm: a $\mathcal{O}(\sqrt{KT})$ bound valid for any probability set $\mathcal{P}$, a $\mathcal{O}(K\ln^2 T)$ bound when $\mathcal{P}$ is a polytope, and a $\mathcal{O}(1)$ regret bound when $\mathcal{P}$ is a polytope contained in the interior of the simplex.

Denote by $\mathrm{Ext}(\mathcal{P})$ the set of extremal points of $\mathcal{P}$; recall that, by definition, a convex $\mathcal{P}$ is a polytope if and only if $\mathrm{Ext}(\mathcal{P})$ is a finite set.

5.2.1. Setting and description of the algorithm

Define the empirical mean associated with arm a at time t

$$\widehat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{A_s=a\}}\,, \quad \text{with the convention that } \widehat{\mu}_a(t) = 1 \text{ if } N_a(t) = 0\,,$$

and introduce the upper confidence bound

$$U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2\ln t}{\max\big(N_a(t),1\big)}}\;; \tag{5.1}$$

denote by $\underline{U}(t)$ the vector with components $U_a(t)$. Note that in the diversity-preserving setting, we cannot to ensure that arm a be picked even once. Therefore, contrary to the vanilla bandit setting, it is important to handle the case when $N_a(t) = 0$. We thus set a default value for $\widehat{\mu}_a(t)$ when $N_a(t) = 0$ to be 1, the highest reward value in the bounded model. This is also the reason why we put a maximum in the denominator of the upper confidence bound.

The natural extension of the UCB algorithm is to pick $\underline{p}_t$ maximizing the scalar product with the upper confidence vector, among the extremal points of $\mathcal{P}$:

$$\underline{p}_{t+1} \in \operatorname*{argmax}_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} \big\langle \underline{p}, \underline{U}(t) \big\rangle \, .$$

Note that the maximum over $\mathcal{P}$ of the linear functional $\underline{p} \mapsto \big\langle \underline{p}, \underline{U}(t) \big\rangle$ is reached for some $\underline{p}$ in $\mathrm{Ext}(\mathcal{P})$. The requirement that $\underline{p}_t$ be chosen among the extremal points only is made for a technical reason; see the first paragraph of Section 5.2.3.

Algorithm 5.1 Diversity-preserving UCB

for rounds $t = 1, \dots,$ **do**
  Select and play
  $$\underline{p}_t \in \operatorname*{argmax}_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} \big\langle \underline{p}, \underline{U}(t-1) \big\rangle \, ,$$
  with ties broken arbitrarily
  Play the pure action $A_t \sim \underline{p}_t$
  Get and observe the reward $Y_t \sim \nu_{A_t}$
  Update the upper confidence bound vector $\underline{U}(t)$ according to the formula, for all $a \in [K]$
  $$U_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{2\ln(t)}{\max\big(N_a(t), 1\big)}} \tag{5.2}$$
end for

Relationship to vanilla and linear bandits. When $\mathcal{P}$ is the whole simplex, then this algorithm picks the pure action a with maximal index $U_a(t-1)$; this is exactly the UCB algorithm for vanilla multi-armed bandits (as soon as all arms are picked once, which is a technical detail.)

Let us discuss the relation of this algorithm to LinUCB for linear bandits Lattimore and Szepesvári [2020]. In this family of algorithms, the main design principle is to build confidence sets $\mathcal{C}_t$ which contain the true mean-payoff vector with high probability, and to choose an action x_t maximizing the payoffs $\big\langle x, \underline{\mu} \big\rangle$ for $\underline{\mu} \in \mathcal{C}_t$. Different types of confidence sets yield different algorithms, and, generally, tighter confidence sets lead to better performance.

Given observations $Y_1, \dots, Y_t$ associated with choices $\underline{p}_1, \dots, \underline{p}_t$, a typical confidence set for linear bandits would be an ellipsoid centered around an estimate $\widehat{\underline{\mu}}_t^{\mathrm{lin}}$. To construct this set, denote by $\mathbb{X}_t$ the matrix whose rows are $\underline{p}_1^\top, \dots, \underline{p}_t^\top$, and $Y_{1:t} = (Y_1, \dots, Y_t)^\top$ and define

$$V_t^\lambda = \lambda I_d + \mathbb{X}_t^\top \mathbb{X}_t \quad \text{and} \quad \widehat{\underline{\mu}}_t^{\mathrm{lin}} = \big(V_t^\lambda\big)^{-1} \mathbb{X}_t^\top Y_{1:t} \, ,$$

where λ is a small regularization parameter. The following ellipsoid is the precise choice made, e.g., in Abbasi-Yadkori et al. [2011]:

$$\mathcal{C}_t^{\mathrm{lin}} = \left\{ \underline{\mu} \in \mathbb{R}^d \, : \, \Big\langle \underline{\mu} - \widehat{\underline{\mu}}_t^{\mathrm{lin}}, \, V_t^\lambda \big(\underline{\mu} - \widehat{\underline{\mu}}_t^{\mathrm{lin}}\big) \Big\rangle \leqslant \frac{1}{4} \sqrt{d \ln\Big(t\big(1 + t/\lambda\big)\Big) + \lambda^{1/2}} \right\} . \tag{5.3}$$

See the previously mentioned reference for a proof the fact that this is indeed a confidence set, i.e., that it contains the true mean-vector with probability at least $1 - 1/t$.

As discussed before, our diversity-preserving setting can be seen as an extension of linear bandits, with extra information observed. The diversity-preserving UCB algorithm follows the optimism principle, but builds tighter confidence sets for $\underline{\mu}$ thanks to the extra information A_t.

Indeed, it uses the simpler rectangular confidence set coming from Hoeffding's inequality, which treats each coordinate independently

$$\mathcal{C}_t = \left\{ (\mu_1, \dots, \mu_K) \in \mathbb{R}^K \ : \ \forall a \in [K], \quad |\widehat{\mu}_a(t) - \mu_a| \leqslant \sqrt{\frac{2 \ln t}{\max(N_a(t), 1)}} \right\} . \tag{5.4}$$

See Figure 5.1 for a comparison of the resulting confidence sets on some simulated data.

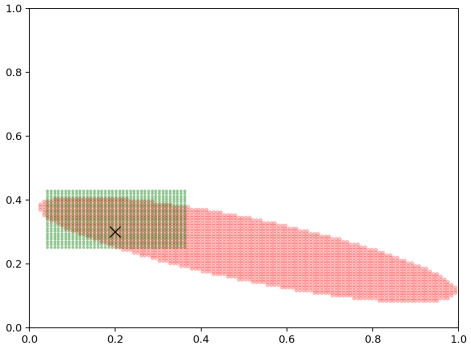


Figure 5.1.: Comparison of confidence sets for Bernoulli observations generated from the three probability vectors $\underline{p}_1 = (0.1, 0.9), \underline{p}_2 = (0.2, 0.8), \underline{p}_3 = (0.4, 0.6)$ and true mean vector $(\mu_1, \mu_2) = (0.2, 0.3)$. Each $\underline{p}_i$ for $i \in \{1, 2, 3\}$ was selected 100 times to generate $A \in \{1, 2\}$, after which reward $Y \sim \mathrm{Ber}(\mu_A)$ was drawn, totalling to $T = 300$ observations. The true mean vector is shown as a blue cross. The red area depicts the ellipsoid defined in (5.3) from the observations, whereas the green rectangle is the one from (5.4).

5.2.2. A distribution-free regret bound

We first provide a distribution-free bound on the regret incurred by our algorithm.

Theorem 5.1. *The diversity-preserving UCB algorithm (Algorithm 5.1) ensures that for all T,*

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}_{[0,1]}} R_T \leqslant 2\sqrt{2 \ln T}\big(K + 2\sqrt{KT}\big) + 2 + K/T^2 \, .$$

Note here that a standard linear bandit algorithm would typically have a guarantee of order $K\sqrt{T}$ (up to multiplicative polylogarithmic terms), which would be worse than our result. More generally, this raises the question of the optimal dependence of the minimax regret on the geometry of $\mathcal{P}$; see Open question 5.4.

Intuituion behind and overview of the proof of Theorem 5.1 . The proof follows quite closely that of Abbasi-Yadkori et al. [2011] for linear bandits, with a few modifications specific to our setting. There are two main steps in the usual proof scheme, which we recall informally. One is to bound the probability that the true mean-vector does not belong to the confidence set at some time step; that probability is small, so we can use a pessimistic bound on the regret incurred in that case.

The other part of the proof consists in handling the regret under the assumption that all the confidence sets indeed contain the mean-vector. Then the difference $\langle \underline{p}^\star - \underline{p}_t, \underline{\mu} \rangle$ between the

payoff of the chosen action and that of the best action can be controlled by a quantity that depends on the size of the confidence set. Thus as the size of the confidence set reduces over time, cumulative regret bounds can be derived.

While we follow this usual proof scheme in our setting, some differences with the analogue linear bandit argument should be mentioned. Of course, the first difference comes from the rectangular shape of the confidence sets, which allows us to treat each coordinate separately, and to bypass the linear algebraic manipulations.

The second difference, perhaps more subtle, lies in the mismatch between the chosen actions and the observations, and in the remaining stochasticity between the choice $\underline{p}_t$ and the observation A_t. This difference manifests itself concretely in the proof at equation (5.6): it is difficult to handle directly the sum

$$\sum_{t=0}^{T-1} \sum_{a=1}^{K} \frac{p_{t+1,a}}{\sqrt{\max\left(N_a(t), 1\right)}},$$

a quantity which involves both the chosen $\underline{p}_t$'s and the $N_a(t)$'s that come from the observations. Indeed, the problem comes from the fact that contrary to the usual linear bandit setting, selecting action $\underline{p}_t$ does not mechanically reduce the size of the confidence set in a predictable way, since various values of A_t's may come from the choice $\underline{p}_t$. Of course the easy fix is to look at this in expectation, as $\mathbb{P}[A_{t+1} = a \,|\, \underline{p}_{t+1}] = p_{t+1,a}$; see the proof for details.

Detailed proof of Theorem 5.1. Let $\underline{p}^\star$ denote an optimal probability in $\mathcal{P}$, and let $r_t = \langle \underline{p}^\star - \underline{p}_t, \underline{\mu} \rangle$ be the instantaneous regret suffered at time t. We begin with an elementary manipulation on r_t. By definition of $\underline{p}_{t+1}$ we have $\langle \underline{p}^\star - \underline{p}_{t+1} \underline{U}(t) \rangle \leqslant 0$ so

$$r_{t+1} = \left\langle \underline{p}^\star - \underline{p}_{t+1}, \underline{\mu} \right\rangle = \langle \underline{p}^\star, \underline{\mu} - \underline{U}(t) \rangle + \underbrace{\left\langle \underline{p}^\star - \underline{p}_{t+1}, \underline{U}(t) \right\rangle}_{\leqslant 0} + \left\langle \underline{p}_{t+1}, \underline{U}(t) - \underline{\mu} \right\rangle .$$

Let us now define a favorable event $\mathcal{E}(t)$ under which the estimates of the means are not too far from the true means. To reduce clutter, we will write $\overline{N}_a(t) = \max(N_a(t), 1)$,

$$\mathcal{E}(t) = \left\{ \text{For all } a \in [K], \quad |\mu_a - \widehat{\mu}_a(t)| \leqslant \sqrt{\frac{2 \ln t}{\overline{N}_a(t)}} \right\} .$$

By concentration, we will see that this event holds with probability at least $1 - 2Kt^{-3}$. Under $\mathcal{E}(t)$, we have for all $a \in [K]$,

$$0 \leqslant U_a(t) - \mu_a \leqslant 2\sqrt{\frac{2 \ln t}{\overline{N}_a(t)}}$$

and therefore, under $\mathcal{E}(t)$, we can further bound the instantaneous regret,

$$r_{t+1} \leqslant \left\langle \underline{p}^\star, \underline{\mu} - \underline{U}(t) \right\rangle + \left\langle \underline{p}_{t+1}, \underline{U}(t) - \underline{\mu} \right\rangle \leqslant 0 + 2 \sum_{a=1}^{K} p_{t+1,a} \sqrt{\frac{2 \ln t}{\overline{N}_a(t)}} . \tag{5.5}$$

Then, by integrating, and using the worst case bound $r_{t+1} \leqslant 1$ when $\mathcal{E}(t)$ does not hold,

$$\mathbb{E}[r_{t+1}] \leqslant \big(1 - \mathbb{P}\{\mathcal{E}(t)\}\big) + 2\mathbb{E}\left[\sum_{a=1}^{K} p_{t+1,a} \sqrt{\frac{2 \ln t}{\overline{N}_a(t)}} \right] \leqslant \big(1 - \mathbb{P}\{\mathcal{E}(t)\}\big) + 2\sqrt{2 \ln T}\, \mathbb{E}\left[\sum_{a=1}^{K} \frac{p_{t+1,a}}{\sqrt{\overline{N}_a(t)}} \right] . \tag{5.6}$$

We also used the fact that $\ln t \leqslant \ln T$. We will deduce an upper bound on the regret by summing over t. Let us start with a rewriting of the second term, thanks to which we will work on the sum. By conditioning on the past observations, i.e., on $\mathcal{F}_{t+1} = \sigma\big(\underline{p}_1, A_1, Y_1, \dots, \underline{p}_t, A_t, Y_t, \underline{p}_{t+1}\big)$, for all $a \in [K]$

$$\mathbb{P}\big[A_{t+1} = a \,\big|\, \mathcal{F}_{t+1}\big] = p_{t+1,a} \quad \text{thus, as } \overline{N}_a(t) \text{ is } \mathcal{F}_{t+1}\text{-measurable,}$$

$$\mathbb{E}\left[\frac{p_{t+1,a}}{\sqrt{\overline{N}_a(t)}}\right] = \mathbb{E}\left[\frac{\mathbb{P}\big[A_{t+1} = a \,\big|\, \mathcal{F}_{t+1}\big]}{\sqrt{\overline{N}_a(t)}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\sqrt{\overline{N}_a(t)}}\,\middle|\, \mathcal{F}_{t+1}\right]\right] = \mathbb{E}\left[\frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\sqrt{\overline{N}_a(t)}}\right].$$

We handle the sum over t thanks to this rewriting. Indeed, for all a, using the fact that $\overline{N}_a(t)$ increases by 1 if and only if $A_{t+1} = a$, and treating

$$\sum_{t=0}^{T-1} \frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\sqrt{\overline{N}_a(t)}} = 1 + \sum_{n=1}^{N_a(T-1)} \frac{1}{\sqrt{n}} \leqslant 1 + 2\sqrt{N_a(T-1)} \leqslant 1 + 2\sqrt{N_a(T)}\,.$$

note that the equality also holds when $N_a(T-1) = 0$, in which case both sums equal 1. Then, by summing the previous inequality over K and by concavity of $\sqrt{\cdot}$ to obtain the second inequality,

$$\sum_{t=0}^{t-1} \sum_{a=1}^{K} \frac{p_{t+1,a}}{\sqrt{\overline{N}_a(t)}} \leqslant K + 2\sum_{a=1}^{K} \sqrt{N_a(T)} \leqslant K + 2\sqrt{KT}\,.$$

Going back to equation (5.6), we have proven that

$$R_T = \sum_{t=1}^{T} r_t \leqslant \sum_{t=0}^{T-1} (1 - \mathbb{P}\big\{\mathcal{E}(t)\big\}) + 2\sqrt{2\ln T}\big(K + 2\sqrt{KT}\big)\,.$$

We now upper bound $\big(1 - \mathbb{P}\big\{\mathcal{E}(t)\big\}\big)$ via a concentration result, stated as Lemma 5.1. As the lemma is only valid for $t \geqslant 2$, we also bound the terms corresponding to $t = 0$ and $t = 1$ by 1. The claimed result follows after some more elementary calculations:

$$R_T \leqslant 2 + \sum_{t=2}^{T-1} 2Kt^{-3} + 2\sqrt{2\ln T}\big(K + 2\sqrt{KT}\big) \leqslant 2 + 2K\frac{1}{2T^2} + 2\sqrt{2\ln T}\big(K + 2\sqrt{KT}\big)$$

which is the stated bound. □

Lemma 5.1. *For $t \geqslant 2$, if the rewards $Y_1, \dots, Y_t$ and actions $A_1, \dots, A_t$ are generated according to Protocol 5.1, and under the bounded problem $(\nu_1, \dots, \nu_K) \in \mathcal{D}_{[0,1]}$ then*

$$\mathbb{P}\left\{\textit{For all } a \in [K], \quad |\mu_a - \widehat{\mu}_a(t)| \leqslant \sqrt{\frac{2\ln t}{\max\big(N_a(t), 1\big)}}\right\} \geqslant 1 - 2Kt^{-3}\,.$$

Proof. By optional skipping (see Section 2.4.1 in Chapter 2), we can replace the random quantities depending on the observations from a fixed arm by their i.i.d. analogue. More precisely, for arm $a \in [K]$ define $\widehat{\mu}_{a,n}$ to be the empirical mean of n i.i.d. samples from ν_a, then

$$\mathbb{P}\left\{|\mu_a - \widehat{\mu}_a(t)| \geqslant \sqrt{\frac{2\ln t}{\overline{N}_a(t)}}\right\} \leqslant \mathbb{P}\left\{\exists n \in [t] \;:\; |\mu_a - \widehat{\mu}_{a,n}| \geqslant \sqrt{\frac{2\ln t}{\max(n,1)}}\right\}$$

$$\leqslant \sum_{n=1}^{t} \mathbb{P}\left\{|\mu_a - \widehat{\mu}_{a,n}| \geqslant \sqrt{\frac{2\ln t}{\max(n,1)}}\right\} \leqslant \sum_{n=1}^{t} 2t^{-4} = 2t^{-3}$$

where the second inequality is a union bound over the value of $N_a(t) \in [t]$, and the third inequality is Hoeffding's inequality. Note that if $n = 0$ and $t \geqslant 2$, since we defined $\widehat{\mu}_{a,0}$ to be 1, this event amounts to $|1 - \mu_a| \geqslant \sqrt{2 \ln t} \geqslant \sqrt{2 \ln 2} > 1$, which cannot happen; thus the bound remains valid. The claimed inequality follows from a union bound over $a \in [K]$. □

5.2.3. A distribution-dependent regret bound when $\mathcal{P}$ is a polytope

As in the previous section, we follow and adapt the proof scheme for the linear bandit algorithms analysis to derive polylogarithmic distribution-dependent regret upper bounds for Algorithm 5.1. Let us sketch this proof.

If $\mathcal{P}$ is finite or is a polytope, then the set of its extremal points $\mathrm{Ext}(\mathcal{P})$ is finite. In that case, the player knows that at least one action in $\mathrm{Ext}(\mathcal{P})$ is optimal; therefore it is a reasonable choice to play only among the extremal points of $\mathcal{P}$, as we have requested in the definition of Algorithm 5.1.

Doing so, the player ensures that the instantaneous regret suffered from playing $\underline{p}$ among the actions in $\mathrm{Ext}(\mathcal{P})$ is either 0, if $\underline{p}$ is optimal, or at least

$$\Delta \stackrel{\text{def}}{=} \min\{\, \Delta(\underline{p}) : \underline{p} \in \mathrm{Ext}(\mathcal{P}),\ \Delta(\underline{p}) > 0 \,\},$$

if $\underline{p}$ is suboptimal. This simple observation leads to the following crude upper bound on the cumulative regret. Denote by $r_t = \langle \underline{p}^\star - \underline{p}_t \,, \underline{\mu} \rangle$ the instantaneous regret suffered at round t, then

$$r_t \leqslant \frac{r_t^2}{\Delta} \quad \text{thus} \quad R_T = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \leqslant \mathbb{E}\left[\sum_{t=1}^{T} \frac{r_t^2}{\Delta}\right] = \frac{1}{\Delta}\mathbb{E}\left[\sum_{t=1}^{T} r_t^2\right].$$

From there on, although we deal with r_t^2 instead of r_t, the analysis is very similar to the distribution-free one. In particular, we define the same favorable event under which all the confidence sets contain the true mean $\underline{\mu}$, and handle the regret under that event.

Theorem 5.2. *If $\mathcal{P}$ is either finite or a polytope, then Algorithm 5.1 enjoys the following distribution-dependent regret bound on any bounded bandit problem in $\mathcal{D}_{[0,1]}$:*

$$R_T \leqslant \frac{8K(\ln T)\big(2 + \ln(T/K)\big) + 2 + K/T^2}{\Delta},$$

where Δ is the minimum gap among suboptimal probabilities in $\mathrm{Ext}(\mathcal{P})$.

This bound is of order $K\big(\ln(T)\big)^2/\Delta$; we feel the time-dependence could be improved, see the following open question.

Open question 5.1. *When $\mathcal{P}$ is finite or is a polytope, a natural strategy is to select probabilities $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ according to the standard UCB algorithm and neglect all the structure of the problem; this strategy yields a regret bound of*

$$\sum_{\underline{p} \in \mathrm{Ext}(\mathcal{P})} 8\frac{\ln T}{\Delta(\underline{p})} + \mathcal{O}\big(\,|\mathrm{Ext}(\mathcal{P})|\,\big)\,,$$

While this approach is definitely not reasonable, it has the theoretical advantage of yielding a problem-dependent regret bound of order $\ln T$, which is better than our guarantee of order $8K(\ln T)^2/\Delta\big(1 + o(1)\big)$. Of course the dependence on other parameters is considerably worse; in particular the bound scales badly with the number of vertices. See Section 5.1.1 for a discussion on the number of vertices.

We feel that it should be possible to derive logarithmic regret bounds with reasonable dependence on all parameters, either for our diversity-preserving UCB or for another algorithm, but we were unable to do so for the time being.

Proof. As explained in the paragraph preceding the statement of the theorem, we start from the upper bound

$$R_T \leqslant \frac{1}{\Delta}\mathbb{E}\left[\sum_{t=1}^{T} r_t^2\right].$$

and define the favorable event

$$\mathcal{E}(t) = \left\{\text{For all } a \in [K], \quad |\mu_a - \widehat{\mu}_a(t)| \leqslant \sqrt{\frac{2\ln t}{\overline{N}_a(t)}}\right\}.$$

Then, once again, under $\mathcal{E}(t)$, we have for all arms $a \in [K]$,

$$0 \leqslant U_a(t) - \mu_a \leqslant 2\sqrt{\frac{2\ln t}{\overline{N}_a(t)}},$$

and, under $\mathcal{E}(t)$, we can further bound the instantaneous regret,

$$r_{t+1} \leqslant \big\langle \underline{p}^\star, \underline{\mu} - \underline{U}(t)\big\rangle + \Big\langle \underline{p}_{t+1}, \underline{U}(t) - \underline{\mu}\Big\rangle \leqslant 0 + 2\sum_{a=1}^{K} p_{t+1,a}\sqrt{\frac{2\ln t}{\overline{N}_a(t)}}.$$

We go one step further by squaring the inequality and applying the Cauchy-Schwarz inequality

$$r_{t+1}^2 \leqslant 4\left(\sum_{a=1}^{K} p_{t+1,a}\sqrt{\frac{2\ln t}{\overline{N}_a(t)}}\right)^2 \leqslant 4\left(\sum_{a=1}^{K}\frac{p_{t+1,a}}{\overline{N}_a(t)}\right)\left(\sum_{a=1}^{K} p_{t+1,a}(2\ln t)\right) = 8\left(\sum_{a=1}^{K}\frac{p_{t+1,a}}{\overline{N}_a(t)}\right)\ln t\,.$$

Thus we have reached an inequality very close to (5.6) in the distribution-free bound:

$$\mathbb{E}[r_{t+1}^2] \leqslant 1 - \mathbb{P}\{\mathcal{E}(t)\} + 8(\ln t)\mathbb{E}\left[\sum_{a=1}^{K}\frac{p_{t+1,a}}{\overline{N}_a(t)}\right] \leqslant 1 - \mathbb{P}\{\mathcal{E}(t)\} + 8(\ln T)\mathbb{E}\left[\sum_{a=1}^{K}\frac{p_{t+1,a}}{\overline{N}_a(t)}\right]. \quad (5.7)$$

The end follows the exact same steps, except that the new bound involves $\overline{N}_a(t)$ instead of $\sqrt{\overline{N}_a(t)}$ in (5.6). Thus, after conditioning by observations up to time t

$$\mathbb{E}\left[\sum_{a=1}^{K}\frac{p_{t+1,a}}{\overline{N}_a(t)}\right] = \mathbb{E}\left[\sum_{a=1}^{K}\frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\overline{N}_a(t)}\right],$$

and, in this order, by rearranging the terms in the sum, by using the fact that the N-th term in the harmonic series is upper bounded by $1 + \ln N$, and by concavity of the logarithm,

$$\sum_{a=1}^{K}\sum_{t=1}^{T}\frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\overline{N}_a(t)} \leqslant 2K + \sum_{a=1}^{K}\sum_{n=1}^{N_a(T-1)}\frac{1}{n} \leqslant 2K + \sum_{a=1}^{K}\ln\big(N_a(T)\big) \leqslant K + K\ln(T/K)\,.$$

Moreover, by Lemma 5.1, we can upper bound $1 - \mathbb{P}\{\mathcal{E}(t)\} \leqslant 2Kt^{-3}$ therefore, going back to (5.7), we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1} r_{t+1}^2\right] &\leqslant \sum_{t=0}^{T-1}\big(1 - \mathbb{P}\{\mathcal{E}(t)\}\big) + 8\ln T\big(K + K\ln(T/K)\big)\\ &\leqslant 2 + \sum_{t=2}^{T} 2Kt^{-3} + 8\ln T\big(K + K\ln(T/K)\big) \leqslant 2 + 2K\frac{1}{2T^2} + 8(\ln T)\big(2K + K\ln(T/K)\big)\end{aligned}$$

which concludes the proof. □

5.2.4. Improving the guarantees

An alternative proof scheme to show that diversity-preserving MOSS has $\sqrt{KT}$ regret. In the vanilla bandit setting, the MOSS algorithm from Audibert and Bubeck [2010] uses a modified version of the UCB index to improve the distribution-free regret guarantees; the index at hand, when adapted to our setting reads

$$U_a^{\mathrm{M}}(t) = \widehat{\mu}_a(t) + \sqrt{\frac{1}{N_a(t)} \ln_+\left(\frac{t}{KN_a(t)}\right)},$$

where $\ln_+(x) = \max(0, \ln x)$, and, say, $U_a^{\mathrm{M}}(t) = 1$ if $N_a(t) = 0$.

We give a sketch of how we could use a different proof scheme to obtain a $O\big(\sqrt{KT}\big)$ regret bound for the MOSS index version of diversity-preserving UCB, shaving off a $\sqrt{\ln T}$.

Instead of following the linear bandit algorithms analysis, we propose to follow the analysis of the MOSS algorithm described in Chapter 2. For the sake of conciseness, we only sketch the main elements of the analysis as most of the steps are exactly treated as in Chapter 2.

Theorem 5.3. *There exists a numerical constant $c > 0$ for which the diversity-preserving MOSS algorithm ensures that for all T,*

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}_{[0,1]}} R_T \leqslant c\sqrt{KT}\,.$$

Proof sketch. By definition of the algorithm, $\big\langle \underline{p}_t, \underline{U}^{\mathrm{M}}(t-1)\big\rangle \geqslant \big\langle \underline{p}^\star, \underline{U}^{\mathrm{M}}(t-1)\big\rangle$ and we can upper bound the instantaneous regret as follows:

$$r_t = \big\langle \underline{p}^\star - \underline{p}_t, \underline{\mu}\big\rangle \leqslant \big\langle p^\star, \underline{\mu} - \underline{U}^{\mathrm{M}}(t-1)\big\rangle + \big\langle \underline{p}_t, \underline{U}^{\mathrm{M}}(t-1) - \underline{\mu}\big\rangle\,,$$

and bound each of these terms separately. Let us inspect this first term:

$$\mathbb{E}\big[\,\big\langle p^\star, \underline{\mu} - \underline{U}^{\mathrm{M}}(t-1)\big\rangle\,\big] = \sum_{a=1}^K p_a^\star\, \mathbb{E}\big[\,\mu_a - U_a^{\mathrm{M}}(t-1)\,\big]\,.$$

It turns out this term is almost exactly treated in the analysis of the MOSS algorithm in Chapter 2, Proposition 2.6 and it is smaller than $c\sqrt{K/t}$ for some constant c; this sums over $t \in [T]$ to a term of order $\sqrt{KT}$. Note in that in Remark 4 accompanying the proof, we observed that the inequality holds regardless of how the arms are picked.

For the second term, we use again a conditioning argument: since $A_t \sim \underline{p}_t$ given the history up to time t,

$$\begin{aligned}\mathbb{E}\Big[\big\langle \underline{p}_t, \underline{U}^{\mathrm{M}}(t-1) - \underline{\mu}\big\rangle\Big] &= \mathbb{E}\left[\sum_{a=1}^K p_{t,a}\big(U_a^{\mathrm{M}}(t-1) - \mu_a\big)\right] \\ &= \mathbb{E}\left[\sum_{a=1}^K \mathbb{1}_{\{A_t=a\}}\big(U_a^{\mathrm{M}}(t-1) - \mu_a\big)\right] = \mathbb{E}\big[U_{A_t}^{\mathrm{M}}(t-1) - \mu_{A_t}\big]\,.\end{aligned}$$

We end up with a term which also appeared in the MOSS analysis. From the said analysis, we can extract a bound

$$\sum_{t=1}^T \mathbb{E}\big[U_{A_t}^{\mathrm{M}}(t-1) - \mu_{A_t}\big] \leqslant c\sqrt{KT}\,.$$

Once again we gloss over some details: we should verify that the proof holds even though the algorithm generating the sequence A_t is not the same, and that the boundary case when $N_a(t) = 0$ for some a does not lead to problems. The total regret bound is of order $\sqrt{KT}$, as claimed. □

We do not know yet whether the MOSS version of this algorithm could deliver improved distribution-dependent regret bounds. We expect that it should yield a regret bound of order $R_T \leqslant \mathcal{O}\big((\ln T)^2/\Delta\big)$ instead of the $K(\ln T)^2/\Delta$, getting rid of a superfluous K term. These would be minor improvements, as we eventually hope to get $\mathcal{O}\big((\ln T)\big)$ regret bounds, see Open question 5.1.

About high-probability (pseudo-)regret bounds In this paragraph, we give a sketch of a strategy to improve our expected regret bounds into high-probability guarantees. Fix a confidence level $\delta > 0$. A careful look at the proof of Theorem 5.1, e.g., at equation (5.5), suggests that we are close to obtaining an upper bound of the form

$$\tilde{R}_T \overset{\text{def}}{=} \sum_{t=1}^{T} \big\langle \underline{p}^\star - \underline{p}_t, \underline{\mu} \big\rangle \leqslant \sum_{t=0}^{T-1} \sum_{a=1}^{K} \frac{\mathbb{1}_{\{A_{t+1}=a\}}}{\sqrt{\overline{N}_a(t)}}\,, \tag{5.8}$$

with high probability (note that we have defined a specific notion of regret here). From such a bound, we could easily deduce a high-probability bound through an application of Berstein's inequality for martingales (Reminder 5.2), which would add an $\mathcal{O}(\sqrt{T\ln(1/\delta)})$ additive factor. However, we have not quite shown (5.8) : the bound (5.5) holds for each t with a probability at least $1 - Kt^{-3}$. These probabilities need to be summed over $t \in [T]$ if we want the global bound on (5.8). Therefore, each of the bounds on r_t should rather hold with proba $1 - \delta/T$. This would require modifying the indices in the diversity-preserving UCB strategy into

$$U_a^{(\delta)}(t) = \widehat{\mu}_a(t) + \sqrt{2\,\frac{\ln\big(1/\delta\big)}{\overline{N}_a(t)}}\,.$$

There is however an important caveat with this modification, which is common in linear bandits: it requires fixing a confidence level in advance, forcing the player to give up, e.g., on anytime strategies.

5.2.5. Bounded regret when $\mathcal{P}$ is in the interior of the simplex

A more surprising result on the diversity-preserving problem is that bounded regret is possible in some circumstances. Indeed, we show in this section that if $\mathcal{P}$ is a polytope (or it is finite), and if $\mathcal{P}$ is strictly included in the interior of the simplex, then the regret of the diversity-preserving UCB algorithm stays finite as $T \to \infty$.

Let us define the minimal probability of choosing any action $a \in [K]$ over all probabilities $\underline{p} \in \mathcal{P}$,

$$p_{\min}(\mathcal{P}) \overset{\text{def}}{=} \min_{\underline{p}\in\mathcal{P}} \min_{a\in[K]} p_a\,.$$

To reduce clutter we will omit the dependence on $\mathcal{P}$ and write $p_{\min}$ in what follows. The set $\mathcal{P}$ is strictly included in the interior of the simplex if and only if $p_{\min} > 0$. When this holds, the player ends up receiving information on any pure action $a \in [K]$ at every time step with probability at least $p_{\min}$, no matter which exact $\underline{p}_t$ she picks.

This phenomenon is reminiscent of what Hao et al. [2020] call *natural exploration* in linear contextual bandits. In their setting, the player may observe information about suboptimal actions even when playing the best action. Therefore, the regime of observations becomes similar to a full-monitoring one, in which the player would observe the reward of every action at every time step, making bounded regret possible. See also Degenne et al. [2018], who describe another similar setting in which extra-observations allow bounded regret.

Recall that when the probability set $\mathcal{P}$ is a polytope, we denote by Δ the minimum gap among suboptimal probabilities in the extremal points $\mathrm{Ext}(\mathcal{P})$.

Theorem 5.4. *If the probability set $\mathcal{P}$ is a polytope, or if it finite, and if $p_{\min} > 0$ the regret of Algorithm 5.1 on any bandit problem with minimal gap Δ is bounded by*

$$R_T \leqslant \frac{24K}{\Delta}\left[\ln\left(\frac{16}{p_{\min}\Delta^2}\ln\left(\frac{16}{p_{\min}\Delta^2}\right)\right)\right]^2 + \frac{3}{\Delta} + 4K + 3K\frac{\mathrm{e}^{-8p_{\min}/\Delta^2}}{p_{\min}^2}\,.$$

Our proof follows naturally the intuition given above. Under our assumption, the number of times every pure action a gets selected, $N_a(t)$, grows linearly with t with high probability. Therefore, the upper confidence terms $\sqrt{(\ln t)/N_a(t)}$ will mechanically reduce over time, even if the player keeps playing the same probability. Eventually, the algorithm will end up discarding all suboptimal probabilities, and stop incurring regret.

Proof. Let $\underline{p}^\star \in \mathrm{Ext}(\mathcal{P})$ be an optimal probability and define the favorable events

$$\mathcal{E}(t) = \left\{\text{For all } a \in [K], \quad |\mu_a - \widehat{\mu}_a(t)| \leqslant \sqrt{\frac{2\ln t}{\max\bigl(N_a(t),1\bigr)}}\right\}$$

and

$$\mathcal{E}'(t) = \left\{\text{For all } a \in [K], \quad \sqrt{\frac{2\ln t}{\max\bigl(N_a(t),1\bigr)}} < \frac{\Delta}{2}\right\}. \tag{5.9}$$

As in the previous proofs, $\mathcal{E}(t)$ is an event under which the confidence set used by the algorithm contains the true mean vector $\underline{\mu}$. In the second event $\mathcal{E}'(t)$, the width of the confidence bands are small enough to discriminate the suboptimal probabilities. Indeed, when both $\mathcal{E}(t)$ and $\mathcal{E}'(t)$ hold, for any suboptimal $\underline{p}$, we have the following chain of inequalities,

$$\begin{aligned}\bigl\langle \underline{U}(t),\, \underline{p}\bigr\rangle = \bigl\langle \widehat{\mu}(t),\, \underline{p}\bigr\rangle + \sum_{a=1}^K p_a\sqrt{\frac{2\ln t}{\max(N_a(t),1)}} &\leqslant \bigl\langle \underline{\mu},\, \underline{p}\bigr\rangle + 2\sum_{a=1}^K p_a\sqrt{\frac{2\ln t}{\max(N_a(t),1)}} \\ &< \bigl\langle \underline{\mu},\, \underline{p}\bigr\rangle + \Delta(\underline{p}) = \bigl\langle \underline{\mu},\, \underline{p}^\star\bigr\rangle \leqslant \bigl\langle \underline{\widehat{\mu}}(t),\, \underline{p}^\star\bigr\rangle + \sum_{a=1}^K p_a^\star\sqrt{\frac{2\ln t}{\max(N_a(t),1)}} = \bigl\langle \underline{U}(t),\, \underline{p}^\star\bigr\rangle\,.\end{aligned}$$

The first inequality follows from the definitions of $\mathcal{E}(t)$ and the second from that of $\mathcal{E}'(t)$. The third inequality is a second application of the definition of $\mathcal{E}(t)$.

We will now show that for times t large enough, $\mathcal{E}(t)$ and $\mathcal{E}'(t)$ will both hold with high probability, ensuring that only optimal probabilities will be picked. As we have seen previously, Lemma 5.1 ensures that for $t \geqslant 2$,

$$\mathbb{P}[\mathcal{E}(t)] \geqslant 1 - 2Kt^{-3}\,.$$

For $\mathcal{E}'(t)$, let us define a time threshold after which $\mathcal{E}'(t)$ becomes likely to happen

$$t_0 \stackrel{\text{def}}{=} \max\bigl\{t \in \mathbb{N} \;\mid\; 8(\ln t)/\Delta^2 > t\,p_{\min}/2\bigr\}\,. \tag{5.10}$$

Then for all $t \geqslant t_0 + 1$, we have $8\ln t/\Delta^2 \leqslant tp_{\min}/2$, and the event $\mathcal{E}'(t)$ will hold if every pure action a gets picked a linear amount of times:

$$\mathbb{P}[\mathcal{E}'(t)] = \mathbb{P}\left\{\text{For all}\, a \in [K], \quad N_a(t) \geqslant \frac{8\ln t}{\Delta^2}\right\} \geqslant \mathbb{P}\left\{\text{For all}\, a \in [K], \quad N_a(t) \geqslant \frac{t\,p_{\min}}{2}\right\}.$$

But of course, every action a is selected at each time step t with probability at least $p_{\min}$, so $N_a(t)$ is likely to grow at least as fast as $tp_{\min}/2$. Formally, the process $(N_a(t) - tp_{\min})$ is a sub-martingale with respect to the filtration $\mathcal{F}_t = \sigma(A_1, Y_1, \dots, A_t, Y_t)$ as,

$$\mathbb{E}\big[N_a(t+1) \,|\, \mathcal{F}_t\big] = N_a(t) + p_{t+1,a} \geqslant N_a(t) + p_{\min} \quad \text{so} \quad \mathbb{E}\big[N_a(t+1) - (t+1)p_{\min} \,|\, \mathcal{F}_t\big] \geqslant N_a(t) - tp_{\min}\,.$$

Moreover, the increments of this sub-martingale are bounded by 1. Therefore, by applying the Azuma-Hoeffding inequality, for any $a \in [K]$, we get

$$\mathbb{P}\left\{ N_a(t) \leqslant \frac{t\,p_{\min}}{2} \right\} = \mathbb{P}\left\{ N_a(t) - t\,p_{\min} \leqslant -\frac{t\,p_{\min}}{2} \right\} \leqslant \exp\left(-\frac{2}{t} \left(\frac{tp_{\min}}{2} \right)^2 \right).$$

Thus for $t \geqslant t_0 + 1$, with a union bound over $a \in [K]$,

$$\mathbb{P}[\mathcal{E}'(t)] \geqslant 1 - K\mathrm{e}^{-tp_{\min}^2/2}\,.$$

Now, since $\Delta(\underline{p}_t) \leqslant 1$ a.s., and since we showed that a suboptimal $\underline{p}_t$ can only be played if $\mathcal{E}(t)$ and $\mathcal{E}'(t)$ do not simultanesouly hold, we have

$$\mathbb{E}[\Delta(\underline{p}_t)] \leqslant \mathbb{P}\big[\mathrm{not}\big(\mathcal{E}(t) \text{ and } \mathcal{E}'(t)\big)\big] \leqslant \mathbb{P}[\text{not } \mathcal{E}(t)] + \mathbb{P}[\text{not } \mathcal{E}'(t)] \leqslant 2Kt^{-3} + K\mathrm{e}^{-tp_{\min}^2/2}\,.$$

Hence the regret can be bounded as

$$R_T = R_{t_0} + \sum_{t=t_0+1}^{T} \mathbb{E}[\Delta(\underline{p}_t)] \leqslant R_{t_0} + 2K \sum_{t=t_0+1}^{T} t^{-3} + K \sum_{t=t_0+1}^{T} \mathrm{e}^{-tp_{\min}^2/2} \leqslant R_{t_0} + 4K + K\frac{\mathrm{e}^{-t_0 p_{\min}^2/2}}{1 - \mathrm{e}^{-p_{\min}^2/2}}\,, \tag{5.11}$$

proving the claim that the regret is finite when $T \to \infty$.

Let us now make this bound more explicit, by computing t_0 and bounding R_{t_0}. At this point, the trivial way to proceed would be to bound R_{t_0} by t_0, but we can easily improve on this by appealling to the earlier analysis of the algorithm. Indeed, by the distribution-dependent bound of Theorem 5.2, if $t_0 \geqslant K$,

$$R_{t_0} \leqslant 24\frac{K(\ln t_0)^2}{\Delta} + \frac{3}{\Delta}\,.$$

Actually, this also holds if $t_0 \leqslant K$, since in that case $R_{t_0} \leqslant t_0 \leqslant K$ and K is smaller than the mentioned bound.

From the definition (5.10) of t_0, we can approximate its value (postponing these calculations to after the proof),

$$\frac{16}{p_{\min}\Delta^2} \leqslant t_0 \leqslant \frac{32}{p_{\min}\Delta^2} \ln\left(\frac{16}{p_{\min}\Delta^2} \right). \tag{5.12}$$

Next, since the function $\varphi : x \mapsto x/(1 - e^{-x})$ is increasing and $0 \leqslant p_{\min}^2/2 \leqslant 1/2$,

$$\frac{1}{1 - \mathrm{e}^{-p_{\min}^2/2}} \leqslant \varphi(1/2)\frac{2}{p_{\min}^2} \leqslant \frac{3}{p_{\min}^2}\,,$$

and using the first inequality in the approximation of t_0 in (5.12), we bound $\mathrm{e}^{-t_0 p_{\min}^2/2} \leqslant \mathrm{e}^{-8p_{\min}/\Delta^2}$. Combining previous results, we have proved the following bound on the regret of diversity-preserving UCB

$$R_T \leqslant \frac{24K}{\Delta}\left[\ln\left(\frac{32}{p_{\min}\Delta^2} \ln\Big(\frac{16}{p_{\min}\Delta^2} \Big) \right) \right]^2 + \frac{3}{\Delta} + 4K + 3K\frac{\mathrm{e}^{-8p_{\min}/\Delta^2}}{p_{\min}^2}$$

which is the claimed statement. □

Proof of (5.12). Fix $a < 1/\mathrm{e}$, and let us show that the threshold $x_0 = \max\{x \in \mathbb{N} \mid \ln x \geqslant ax\}$ satisfies

$$\frac{1}{a} \leqslant x_0 \leqslant \frac{2}{a}\ln\left(\frac{1}{a}\right).$$

The approximation (5.12) of t_0 derives from this applied to $a = p_{\min}\Delta^2/16$, which is indeed less than $1/e$.

The first inequality comes from the fact that the mapping $\psi : x \mapsto ax - \ln x$ is increasing on $(1/a, +\infty)$ and $\psi(1/a) = 1 + \ln a < 0$ since we assumed that $a < 1/e$. The second inequality is verified directly as

$$\ln\left(\frac{2}{a}\ln\left(\frac{1}{a}\right)\right) = \ln\frac{1}{a} + \underbrace{\ln\left(2\ln\left(\frac{1}{a}\right)\right)}_{\leqslant \ln(1/a)} < 2\ln\frac{1}{a} = a\left(\frac{2}{a}\ln\left(\frac{1}{a}\right)\right),$$

where we used the fact that $2\ln(x) < x$, which holds for all $x > 0$. □

Open question 5.2. *Is the upper bound on the constant regret close to optimal? While there is certainly room for improvement in this bound, one may wonder whether the $\mathcal{O}(1/\Delta)$ dependence on Δ is improvable.*

Open question 5.3. *We conjecture that we would still get bounded regret under the less restrictive condition that only the optimal actions lie in the interior of the simplex, i.e.,*

$$\min_{\underline{p}^\star \in \mathrm{Opt}(\underline{\nu})} \min_{a \in [K]} p^\star_a > 0\,.$$

Intuitively, diversity-preserving UCB should start selecting optimal probabilities most of the time, as we already know that it enjoy sub-linear regret bounds. Thus the same reasoning as above should apply with this new assumption. This simple-looking conjecture turned out to be harder to prove than we anticipated.

The main reason for this difficulty is that using some prior regret bounds would only control in expectation the numbers $N_a(t)$ of pulls of suboptimal arms. In order to guarantee that the confidence bounds decrease, and extend the proof, we would rather need lower bounds on these variables with high probability.

5.3. A follow-the-regularized-leader approach

In this section, we describe how the diversity-preserving setting can be extended to the adversarial case. In this modified setting, the rewards associated with an arm $a \in [K]$ are chosen arbitrarily by the environement, and are not necessarily i.i.d. The follow-the-regularized-leader (FTRL) algorithm can be used here, and easily provides $\mathcal{O}(\sqrt{KT\ln K})$ regret bound. We also discuss some improvements when the probability set $\mathcal{P}$ is included in the interior of the simplex.

Recall that $\mathcal{S}$ denotes the set of all probability distributions over $[0, 1]$.

5.3.1. (Oblivious) adversarial setting

A set $\mathcal{P}$ of diversity-preserving distributions is fixed. An adversary chooses beforehand a sequence of vectors $\underline{y}_1, \ldots, \underline{y}_t, \ldots$ with coordinates in $[0, 1]$. At every time step t, the player selects a probability vector $\underline{p}_t \in \mathcal{P}$, and samples $A_t \sim \underline{p}_t$. She then observes and receives the reward y_{t,A_t}. The natural measure of performance in this setting is the regret

$$\max_{\underline{p} \in \mathcal{P}} \sum_{t=1}^{T} \langle \underline{p}, \underline{y}_t \rangle - \sum_{t=1}^{T} y_{t,A_t}\,,$$

and its expectation equals

$$R_T(\underline{y}_{1:T}) = \mathbb{E}\left[\max_{\underline{p}\in\mathcal{P}} \sum_{t=1}^{T} \langle \underline{p}, \underline{y}_t\rangle - \sum_{t=1}^{T} y_{t,A_t}\right] = \max_{\underline{p}\in\mathcal{P}} \sum_{t=1}^{T} \langle \underline{p}, \underline{y}_t\rangle - \mathbb{E}\left[\sum_{t=1}^{T} y_{t,A_t}\right] .$$

Note that the equality holds because we assumed that the reward vectors $\underline{y}_t$ are deterministic and chosen before the game is played (the adversary is said to be oblivious, as opposed to the general reactive case in which the adversary can adapt the reward vectors to the player's choices).

5.3.2. Follow-the-regularized-leader algorithms

FTRL is a rich family of online algorithms; see Section 4.C of Chapter 4 a for more thorough presentation, including a literature review. We present a specific variant involving the negentropy as the regularizer, which is defined as

$$H(\underline{p}) = \sum_{a=1}^{K} p_a \ln p_a \,.$$

Using this regulariser, the FTRL approach consists, in our setting, in choosing the mixed action

$$\underline{p}_t \in \operatorname*{argmax}_{\underline{p}\in\mathcal{P}} \left\{ \sum_{s=1}^{t-1} \langle \underline{p}, \widehat{\underline{y}}_s\rangle - \frac{H(\underline{p})}{\eta} \right\}, \tag{5.13}$$

where $\widehat{\underline{y}}_s$ are estimates of the unobserved reward vectors $\underline{y}_s$ built with the observations and η is a fixed learning rate. In the rest of this section, we use the importance weighting estimator $\widehat{\underline{y}}_t$ with coordinates

$$\widehat{y}_{t,a} = \frac{y_{t,A_t} - 1}{p_{t,a}} \mathbb{1}_{\{A_t = a\}} + 1 \,.$$

These estimates are centered around 1 in order to ensure that $\widehat{y}_{t,a} \leqslant 1$, an important property in the following results. Other useful properties are that for all $a \in [K]$, for all $t \geqslant 1$,

$$\mathbb{E}\big[\widehat{y}_{t,a} \mid \underline{p}_t\big] = y_{t,a} \qquad \text{and} \qquad \langle \underline{p}_t, \widehat{\underline{y}}_t\rangle = y_{t,A_t} \,, \tag{5.14}$$

where the first equality holds since A_t is sampled according to $\underline{p}_t$ and independently from the past.

Computing the updates. When $\mathcal{P} = \mathcal{S}$, it is well-known that FTRL with H as the regularizer is exactly the Exp3 algorithm (see Chapter 4 Section 4.C.2). In other words, when $\mathcal{P} = \mathcal{S}$, the updates can be explicitly computed and their value is

$$\tilde{p}_{a,t} = \exp\left(\eta \sum_{s=1}^{t-1} \widehat{y}_{s,a}\right) \Bigg/ \sum_{j=1}^{K} \exp\left(\eta \sum_{s=1}^{t-1} \widehat{y}_{s,j}\right) .$$

This yields an alternative formulation for the updates in the diversity-preserving setting, using Theorem 26.15 in Lattimore and Szepesvári [2020]. This theorem states that the argmax over $\mathcal{P}$ of the objective function (5.13) (which is Legendre) can be computed in a two-step fashion. First, compute the argmax of the objective over $\mathcal{S}$; this is $\tilde{\underline{p}}_t$. Then, project this probability vector to the constraint set $\mathcal{P}$ according to the Bregman divergence associated with H, which is the Kullback-Leibler divergence, i.e.,

$$\underline{p}_t \in \operatorname*{argmin}_{\underline{p}\in\mathcal{P}} \mathrm{KL}\big(\underline{p}, \tilde{\underline{p}}_t\big) \,.$$

5.3.3. General analysis for the expected regret

We recall the standard full-information regret bound for the FTRL algorithm, formulated directly with the reward estimates $\widehat{\underline{y}}_t$; the next reminder is a concatenation of Reminder 4.5 (the general FTRL bound) and Lemma 4.2 (a specification of the bound for negentropy) in Chapter 4; both results are proved therein.

Reminder 5.1. *For any $\underline{p} \in \mathcal{P}$ and for any reward estimates $\widehat{y}_t$ such that $\widehat{y}_{t,a} \leqslant 1$, playing according to the FTRL updates* (5.13) *ensures that, a.s.,*

$$\sum_{t=1}^{T} \langle \underline{p} - \underline{p}_t, \widehat{\underline{y}}_t \rangle \leqslant \frac{H(\underline{p}) - H(\underline{p}_1)}{\eta} + \sum_{t=1}^{T} \frac{1}{2} \sum_{a=1}^{K} p_{t,a}(1 - \widehat{y}_{t,a})^2 \,. \tag{5.15}$$

From this we instantly obtain a regret bound for FTRL in the diversity-preserving setting. We denote by $D_H(\mathcal{P}) = \max\{H(\underline{p}) - H(\underline{q}) \mid \underline{p}, \underline{q} \in \mathcal{P}\}$ the H diameter of $\mathcal{P}$. Note that we have $D_H(\mathcal{P}) \leqslant D_H(\mathcal{S}) = \ln K$.

Theorem 5.5. *For any reward sequence $\underline{y}_{1:T} \in [0,1]^{KT}$, the FTRL algorithm with negentropy, tuned with any $\eta > 0$, enjoyrs the following bound on the expected regret,*

$$R_T(\underline{y}_{1:T}) \leqslant \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta}{2} KT \,; \tag{5.16}$$

in particular, $R_T(\underline{y}_{1:T}) \leqslant \sqrt{2 D_H(\mathcal{P})} \sqrt{KT}$ with $\eta = \sqrt{2 D_H(\mathcal{P})/(KT)}$.

An interesting feature of the theorem above is that it naturally involves the H–diameter of $\mathcal{P}$, a quantity that depends on the geometry of $\mathcal{P}$; see Open question 5.4 for a discussion on the optimal minimax regret.

Proof. Let $p^\star \in \mathcal{P}$ be an optimal probability, that is, a probabiltity maximizing the scalar product with $\underline{y}_1 + \cdots + \underline{y}_T$; recall that $p^\star$ is deterministic since the rewards $\underline{y}_{1:T}$ are themselves deterministic. Then,

$$R_T(\underline{y}_{1:T}) = \sum_{t=1}^{T} \langle \underline{p}^\star, \underline{y}_t \rangle - \mathbb{E}\left[\sum_{t=1}^{T} y_{t,A_t}\right] = \sum_{t=1}^{T} \mathbb{E}\big[\langle \underline{p}^\star, \widehat{\underline{y}}_t \rangle\big] - \mathbb{E}\left[\sum_{t=1}^{T} \langle \underline{p}_t, \widehat{\underline{y}}_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p^\star - \underline{p}_t, \widehat{\underline{y}}_t \rangle\right],$$

where the first equality holds thanks to (5.14). Furthermore, Reminder (5.1) yields,

$$\sum_{t=1}^{T} \langle \underline{p} - \underline{p}_t, \widehat{\underline{y}}_t \rangle \leqslant \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{a=1}^{K} p_{t,a}(1 - \widehat{y}_{t,a})^2 = \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} p_{t,A_t}^{-1}(1 - y_{t,A_t})^2 \,. \tag{5.17}$$

where the equality comes from substituting the definition of $\widehat{\underline{y}}_t$. Then, integrating the summand in the left-most side, and using the fact that $y_{t,a} \in [0,1]$ for all t and a,

$$\frac{1}{2}\mathbb{E}\big[p_{t,A_t}^{-1}\,(1 - y_{t,A_t})^2\big] = \frac{1}{2} \sum_{a=1}^{K} (1 - y_{t,a})^2 \leqslant \frac{1}{2} K \,.$$

Together with the two previous inequalities, this yields the claimed result. □

5.3.4. A high-probability regret bound when $\mathcal{P}$ is in the interior of the simplex

In the stochastic setting, the diversity-preserving problem is easier when $\mathcal{P}$ is fully in the interior of the simplex. We show a similar phenonemon in the adversarial case, although the sense in which the problem becomes easier is different.

Up to here, all the results presented in this chapter dealt with the expected regret suffered by the player. It can be argued that in practice, the player would rather want an algorithm ensuring large rewards with high probability, i.e., enjoying high-probability regret bounds.

In the standard adversarial bandit setting, high-probability regret bounds require making some heavy adjustments to the usual algorithms. Among other changes, in order to reduce the variance of the reward estimates, most methods require adding some extra-exploration, either in some explicit (e.g., in Auer et al. [2002b]) or implicit (e.g., in Kocák et al. [2014]) form.

The next proposition states that the standard FTRL algorithm with negentropy, with no modifications, enjoys high-probability regret bounds whenever the probability set $\mathcal{P}$ is included in the interior of the simplex. This is because lower bounding the minimal probability naturally reduces the variance of the reward estimates.

Recall that we denote by $p_{\min}$ the minimum of the coordinates of all probability vectors in $\mathcal{P}$. The proof is a straighforward application of Bernstein's inequality for martingales, which we recall in Reminder 5.2 below.

Proposition 5.1. *If $p_{\min} > 0$ then, for any reward sequence $\underline{y}_{1:T}$, with probability at least $1-\delta$,*

$$\max_{p\in\mathcal{P}} \sum_{t=1}^{T} \langle \underline{p}, \underline{y}_t \rangle - y_{t,A_t} \leqslant \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta T}{2p_{\min}} + \sqrt{\frac{2T}{p_{\min}} \ln\frac{1}{\delta}} + \frac{1}{3p_{\min}} \ln\frac{1}{\delta}\,.$$

In particular, if $\eta = \sqrt{2D_H(\mathcal{P})p_{\min}/T}$, we get that with probability at least $1-\delta$,

$$\max_{p\in\mathcal{P}} \sum_{t=1}^{T} \langle \underline{p}, \underline{y}_t \rangle - y_{t,A_t} \leqslant \sqrt{\frac{2D_H(\mathcal{P})T}{p_{\min}}} + \sqrt{\frac{2T}{p_{\min}} \ln\frac{1}{\delta}} + \frac{1}{3p_{\min}} \ln\frac{1}{\delta}\,.$$

Note that $p_{\min} \leqslant 1/K$, so this bound is never smaller than that of Theorem 5.5; the fact that it holds with high probability is an improvement.

Proof. Remembering that $y_{t,A_t} = \langle \underline{p}_t, \widehat{\underline{y}}_t \rangle$, the regret can be decomposed as

$$R_T = \sum_{t=1}^{T} \langle \underline{p}^\star, \underline{y}_t \rangle - \langle \underline{p}_t, \widehat{\underline{y}}_t \rangle = \sum_{t=1}^{T} \langle p^\star, \underline{y}_t - \widehat{\underline{y}}_t \rangle + \sum_{t=1}^{T} \langle \underline{p}^\star - \underline{p}_t, \widehat{\underline{y}}_t \rangle\,.$$

By previous results (equation (5.17)), denoting by $\underline{p}^\star$ an optimal probability,

$$\sum_{t=1}^{T} \langle \underline{p}^\star - \underline{p}_t, \widehat{\underline{y}}_t \rangle \leqslant \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \frac{1}{p_{t,A_t}} \leqslant \frac{D_H(\mathcal{P})}{\eta} + \frac{\eta T}{2p_{\min}}\,. \tag{5.18}$$

Now let us show that the left-hand side of this inequality is not too far from the regret, by using Bernstein's inequality for martingales. To do so, consider the filtration $(\mathcal{F}_t)$ generated by $(A_1, \dots, A_t)$, and define the process

$$\left(\sum_{s=1}^{t} \langle \underline{p}^\star, \underline{y}_s - \widehat{\underline{y}}_s \rangle \right)_{t\geqslant 1}, \text{ which is an } (\mathcal{F}_t)\text{–martingale as } \quad \mathbb{E}\big[\langle \underline{p}^\star, \widehat{\underline{y}}_t \rangle \mid \mathcal{F}_{t-1}\big] = \langle \underline{p}^\star, \underline{y}_t \rangle\,.$$

Indeed the equality holds, since the selected probability $\underline{p}_t$ is $\mathcal{F}_{t-1}$ measurable, and $A_t \sim \underline{p}_t$ is drawn independently from the past. The increments are bounded by above since for all a and t,

$$y_{t,a} - \widehat{y}_{t,a} = y_{t,a} - \left((1 - \frac{1 - y_{t,A_t}}{p_{t,a}}\right)\mathbb{1}_{\{A_t=a\}} \leqslant \frac{1 - y_{t,A_t}}{p_{\min}} \leqslant \frac{1}{p_{\min}}\,,$$

and we bound the conditional variance of the increments as follows,

$$\begin{aligned}\mathbb{E}\left[\langle \underline{p}^\star, \widehat{\underline{y}}_t - \underline{y}_t\rangle^2 \mid \mathcal{F}_{t-1}\right] &\leqslant \mathbb{E}\left[\left(\langle \underline{p}^\star, \widehat{\underline{y}}_t\rangle - 1\right)^2 \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[p^\star_{A_t}\left(\frac{1 - y_{t,A_t}}{p_{t,A_t}}\right)^2 \,\Big|\, \mathcal{F}_{t-1}\right] = \sum_{a=1}^K p_{t,a}\, p^\star_a \frac{(1-y_{t,a})^2}{p^2_{t,a}} \leqslant \frac{1}{p_{\min}}\,,\end{aligned}$$

where the first inequality holds because the conditional expectation $\langle \underline{p}^\star, \underline{y}_t\rangle$ minimizes the functional $c \mapsto \mathbb{E}\big[\big(\langle \underline{p}^\star, \underline{y}_t\rangle - c\big)^2 | \mathcal{F}_{t-1}\big]$. Thus the sum of the conditional variances is less than $T/p_{\min}$. Therefore, applying Bernstein's inequality for martingales, with probability at least $1-\delta$,

$$\sum_{t=1}^T \langle p^\star, \underline{y}_t - \widehat{\underline{y}}_t\rangle \leqslant \sqrt{\frac{2T}{p_{\min}} \ln\frac{1}{\delta}} + \frac{1}{3p_{\min}} \ln\frac{1}{\delta}\,.$$

This yields the claimed bound when combined with (5.18). □

Reminder 5.2 (Bernstein's inequality for martingales). *Let $(X_n)_{n\geqslant 1}$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_n)_{n\geqslant 0}$, and let $N \geqslant 1$ be a summation horizon. Assume that there exist real numbers b and v_N such that, almost surely,*

$$\forall n \leqslant N, \quad X_n \leqslant b \qquad \text{and} \qquad \sum_{n=1}^N \mathbb{E}\big[X_n^2 \,\big|\, \mathcal{F}_{n-1}\big] \leqslant v_N\,.$$

Then for all $\delta \in (0,1)$,

$$\mathbb{P}\left[\sum_{n=1}^N X_n \geqslant \sqrt{2v_N \ln\frac{1}{\delta}} + \frac{b}{3}\ln\frac{1}{\delta}\right] \leqslant \delta\,.$$

5.4. A distribution-free lower bound

In this section we prove a distribution-free lower bound on the regret any algorithm has to incur when playing the diversity-preserving problem. This lower bound is to be compared to the upper bounds of Theorems 5.1, 5.3 and 5.5; it shows in particular that the $\sqrt{KT}$ guarantee is close to optimal for some probability sets $\mathcal{P}$.

To formulate the lower bound, we introduce the quantity

$$M(\mathcal{P}) \stackrel{\text{def}}{=} \frac{1}{K}\sum_{i=1}^K \max_{p\in\mathcal{P}} p_i - \frac{1}{K}\,,$$

which is a measure of the width of the probability set $\mathcal{P}$. Note that $M(\mathcal{P})$ ranges between 0 and $1 - 1/K$. Indeed $M(\mathcal{P}) \geqslant 0$ by sub-additivity of the maximum. Moreover, since M is a non-decreasing quantity for the inclusion order, $M(\mathcal{P})$ is smaller than its value when the probability set is the whole simplex, which is $1 - 1/K$.

Proposition 5.2 states that the worst-case regret of any algorithm grows with $M(\mathcal{P})^2$, and can be as large as $\sqrt{KT}$ when $M(\mathcal{P})^2$ is close to 1.

Proposition 5.2. *Let $\mathcal{P}$ be any diversity-preserving probability set. For any bandit algorithm, and for any $T \geqslant KM(\mathcal{P})^2$,*

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}_{[0,1]}} R_T(\underline{\nu}) \geqslant 0.23\, M(\mathcal{P})^2 \sqrt{KT}\,.$$

The proof is a direct adaptation of the minimax lower bound for standard bandits, see Auer et al. [2002b] and Garivier et al. [2019] for an in-depth discussion on bandit lower bounds. The technique involves considering a well-chosen family of bandit problems that are close to each other in a statistical sense. Then we show that any algorithm must play suboptimal actions before it gathers enough information to find out which problem it is playing against.

For this specific lower bound, we use the same family of problems as the ones considered in the lower bound for standard bandits. These are K–action Bernoulli bandit problems with means $1/2$, except for one action with a slightly larger mean.

Proof. Consider the family of bandit problems $\underline{\nu}^{(0)}, \dots, \underline{\nu}^{(K)}$, defined for $i = 0, \dots, K$, to be Bernoulli distributions with means

$$\mu_a^{(i)} = \frac{1}{2} + \varepsilon \mathbb{1}_{\{a=i\}}\,,$$

where $\varepsilon \in [0, 1/4]$ is a constant whose value we will set later on. Note that this also defines a the problem $\mu^{(0)} = (1/2, \dots, 1/2)$ with equal means. For $i \in \{0, \dots, K\}$, denote by $\mathbb{P}_i$ the law of all the observations for problem i at round T, and $\mathbb{E}_i$ the expectation taken according to $\mathbb{P}_i$.

Let us now rewrite the expression of the regret for the problems considered. For a fixed $i \in [K]$ and for any $\underline{p} = (p_1, \dots, p_K) \in \mathcal{P}$,

$$\big\langle \underline{p},\, \mu^{(i)} \big\rangle = \frac{1}{2} + \varepsilon p_i\,,$$

and therefore,

$$R_T(\underline{\nu}^{(i)}) = \varepsilon \left(T \max_{\underline{p} \in \mathcal{P}} p_i - \mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right] \right) = \varepsilon T \left(\max_{\underline{p} \in \mathcal{P}} p_i - \mathbb{E}_i \left[\frac{1}{T} \sum_{t=1}^{T} p_{t,i} \right] \right).$$

Finally, note that by the tower rule,

$$\mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right] = \mathbb{E}_i \big[N_i(T) \big]\,, \quad \text{so that} \quad R_T(\underline{\nu}^{(i)}) = \varepsilon T \left(\max_{\underline{p} \in \mathcal{P}} p_i - \mathbb{E}_i \left[\frac{N_i(T)}{T} \right] \right).$$

Thus by averaging over $i \in [K]$,

$$\frac{1}{K} \sum_{i=1}^{K} R_T(\underline{\nu}^{(i)}) = \varepsilon T \left(\frac{1}{K} \sum_{i=1}^{K} \max_{\underline{p} \in \mathcal{P}} p_i - \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_i \left[\frac{N_i(T)}{T} \right] \right). \tag{5.19}$$

From there on the proof follows quite closely the one of the bandits distribution-free lower bound. By the chain rule for the Kullback-Leibler divergence, see Section 2.1 in Garivier et al. [2019], and denoting by kl the Bernoulli Kullback-Leibler divergence,

$$\mathrm{KL}\big(\mathbb{P}_0, \mathbb{P}_i\big) = \mathbb{E}_0[N_i(T)]\ \mathrm{kl}\big(1/2, 1/2 + \varepsilon\big) \leqslant \mathbb{E}_0[N_i(T)]\, c_0\, \varepsilon^2\,,$$

where we used the fact that $\mathrm{kl}(1/2, 1/2 + \varepsilon) \leqslant c_0 \varepsilon^2$ with $c_0 \leqslant 2.31$, whenever $0 \leqslant \varepsilon \leqslant 1/4$. Now by Pinsker's inequality,

$$\mathrm{KL}\big(\mathbb{P}_0, \mathbb{P}_i\big) \geqslant \sup_{Z \in [0,1]} 2\Big(\mathbb{E}_0[Z] - \mathbb{E}_1[Z]\Big)^2 \geqslant 2\left(\mathbb{E}_0\left[\frac{N_i(T)}{T}\right] - \mathbb{E}_i\left[\frac{N_i(T)}{T}\right]\right)^2,$$

where the supremum is taken over all random variables Z that are bounded in $[0,1]$ and is specified to the random variable $N_i(T)/T$. Next, taking a square root and dividing by 2,

$$\mathbb{E}_i\left[\frac{N_i(T)}{T}\right] \leqslant \mathbb{E}_0\left[\frac{N_i(T)}{T}\right] + \sqrt{\frac{\mathrm{KL}\left(\mathbb{P}_0, \mathbb{P}_i\right)}{2}}\,.$$

Thus, averaging over $i \in \{1,\dots,K\}$, using the concavity of $x \mapsto \sqrt{x}$, and the fact that the variables $N_i(T)/T$ sum up to 1,

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\left[\frac{N_i(T)}{T}\right] \leqslant \frac{1}{K} + \frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_i)}{2}} \leqslant \frac{1}{K} + \sqrt{\frac{1}{2K}\sum_{i=1}^{K}\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_i)} \leqslant \frac{1}{K} + \varepsilon\sqrt{\frac{c_0 T}{2K}}\,. \tag{5.20}$$

Substituting this into (5.19) yields

$$\frac{1}{K}\sum_{i=1}^{K} R_T(\underline{\nu}^{(i)}) \geqslant \varepsilon T\left(\frac{1}{K}\sum_{i=1}^{K}\max_{p\in\mathcal{P}} p_i - \frac{1}{K} - \varepsilon\sqrt{\frac{c_0 T}{2K}}\right) = \varepsilon T\left(M(\mathcal{P}) - \varepsilon\sqrt{\frac{c_0 T}{2K}}\right)$$

Now take $\varepsilon = M(\mathcal{P})\sqrt{K/T}/2$, which is smaller than $1/4$ by the assumption $T \geqslant K M(\mathcal{P})^2$ (note also that $M(\mathcal{P})$ is always non-negative), to get

$$\frac{1}{K}\sum_{i=1}^{K} R_T(\underline{\nu}^{(i)}) \geqslant \frac{1}{2}\left(1 - \sqrt{\frac{c_0}{8}}\right) M(\mathcal{P})^2\sqrt{KT} \geqslant 0.23\, M(\mathcal{P})^2\sqrt{KT}\,.$$

Therefore at least one problem among $\underline{\nu}^{(1)},\dots,\underline{\nu}^{(K)}$ has a regret lower bounded as above, which proves the claimed statement. □

Open question 5.4 (Optimal distribution-free dependence on the geometry $\mathcal{P}$). *Combining this lower bound with the upper bounds of Theorems 5.1 and 5.5, we have obtained the following inequalities on the minimax regret for a given probability set $\mathcal{P}$:*

$$0.23\,\mathcal{M}(\mathcal{P})\sqrt{KT} \leqslant \sup_{\text{strat.}}\inf_{\underline{\nu}\,\text{in}\,\mathcal{D}} R_T \leqslant \min\left(c\sqrt{KT}, \sqrt{2D_H(\mathcal{P})}\sqrt{KT}\right),$$

where the infimum is taken over all strategies and the supremum, over all bandit problems in $\mathcal{D}$, and c is a numerical constant. This shows that $\sqrt{KT}$ is the optimal dependence for some probability sets: those for which $\mathcal{M}(\mathcal{P})$ is close to 1.

However the gap between the left-most and the right-most sides of the inequalities can be significant for some specific $\mathcal{P}$. We would like to reduce this gap, either by improving the analyses of our current algorithms, or by considering new algorithms. One promising approach is to find other regularizing functions in FTRL, chosen depending of the geometry of $\mathcal{P}$.

At the time of writing, we do not have a clear-cut opinion on whether the optimal dependency on the geometry of $\mathcal{P}$ should be expressed in terms of $\mathcal{M}(\mathcal{P})$. It appears naturally in the proof of the lower bound, but we have no corresponding regret upper bound.

5.5. Distribution-dependent regret lower bound for polytopes

The end goal of this section is to study the case of polytopes $\mathcal{P}$, which are given, by definition, by convex hulls of their extremal sets $\mathrm{Ext}(\mathcal{P})$. We derive an asymptotic lower bound on the regret for the diversity-preserving bandit problem. This lower bound is analogous to the well-known lower bound for structured bandits from Graves and Lai [1997]. It is formulated via an optimization

problem, which we study; we derive in particular a necessary and sufficient condition under which the lower bound is positive.

To do so, we first consider the simpler case of finite sets $\mathcal{P}$ and explain later, in Section 5.5.3, how to extend with no effort the results for finite sets to polytopes.

Thus we fix a finite probability set $\mathcal{P}$. With no loss of generality, we assume that for all $a \in [K]$, there is at least a mixed action $\underline{p} \in \mathcal{P}$ such that $p_a > 0$. Otherwise, we may discard arms not fulfilling this condition. Throughout this section, we also fix an arbitrary bandit model $\mathcal{D}$, that is, a set of reward distributions with finite first moment. We will sometimes denote by $I \subseteq \mathbb{R}$ the range of possible means for distributions in $\mathcal{D}$.

5.5.1. An asymptotic lower bound in the diversity-preserving setting / Finite $\mathcal{P}$

For a bandit problem $\underline{\nu}'$ in $\mathcal{D}$ with means $\underline{\mu}'$, we denote by $\mathrm{Opt}(\underline{\nu}')$ the set of its optimal probabilities:

$$\mathrm{Opt}(\underline{\nu}') = \operatorname*{argmax}_{\underline{p} \in \mathcal{P}} \langle \underline{p}, \underline{\mu}' \rangle \,. \tag{5.21}$$

Consider $\underline{\nu} = (\nu_1, \dots, \nu_K)$ in $\mathcal{D}$ a fixed bandit problem. We will assume in this section that the set of optimal actions for $\underline{\nu}$ is a singleton and denote its unique element $\mathrm{Opt}(\underline{\nu}) = \{\underline{p}^\star(\underline{\nu})\}$. This assumption is common in bandit analyses (it is made, e.g., in Lattimore and Szepesvári [2017], Combes et al. [2017]), and it is arguably harmless as generic problems will typically have a unique optimal action.

We make a uniformity assumptions on the performance of the strategies considered. An algorithm is said to be *uniformly fast convergent* (abbreviated to UFC in what follows) over $\mathcal{D}$ if for any bandit problem $\underline{\nu}'$ in $\mathcal{D}$, its regret decays faster than any power of T, i.e., if $R_T(\underline{\nu}') = o(T^\alpha)$ for all $\alpha > 0$. This assumption is satisfied, e.g., by the diversity-preserving UCB algorithm, by Theorem 5.2.

To formulate the lower bound, let us also introduce the set of confusing alternative problems associated with the bandit problem $\underline{\nu}$, denoted by $\textsc{Alt}(\underline{\nu})$. Problems in $\textsc{Alt}(\underline{\nu})$ are the ones in which $p^\star(\underline{\nu})$ is suboptimal, but that the player cannot discriminate from $\underline{\nu}$ by only playing $p^\star(\underline{\nu})$. Precisely, for each arm a, either $p^\star_a(\underline{\nu}) = 0$ and selecting the optimal probability $\underline{p}^\star(\underline{\nu})$ never results in picking arm a, or $\nu_a = \nu'_a$ and observing a reward associated with a does not provide discriminative information information; formally,

$$\textsc{Alt}(\underline{\nu}) \overset{\text{def}}{=} \left\{ \underline{\nu}' \text{ in } \mathcal{D} \;\middle|\; p^\star(\underline{\nu}) \notin \mathrm{Opt}(\underline{\nu}') \;\text{ and }\; \forall\, 1 \leqslant a \leqslant K, \quad p^\star_a(\underline{\nu}) = 0 \;\text{ or }\; \nu_a = \nu'_a \right\} .$$

Since UCB played over all probabilities in $\mathcal{P}$ yields logarithmic regret, we expect the correct scaling of the suboptimal pulls to be logarithmic. Therefore define the normalized allocations

$$n_t(\underline{p}) = \frac{\mathbb{E}_{\underline{\nu}}\big[N^{(\underline{p})}(t)\big]}{\ln t}, \quad \text{so that} \quad \frac{R_t}{\ln t} = \sum_{\underline{p} \in \mathcal{P}} \Delta(\underline{p}) \frac{\mathbb{E}_{\underline{\nu}}\big[N^{(\underline{p})}(t)\big]}{\ln t} = \sum_{\underline{p} \in \mathcal{P}} \Delta(\underline{p})\, n_t(\underline{p}) \,. \tag{5.22}$$

A UFC algorithm facing the problem $\underline{\nu}$ will eventually focus on the unique optimal mixed action $\underline{p}^\star(\underline{\nu})$. Doing so, most of its observations will correspond to pure actions $a \in [K]$ such that $p_a(\underline{\nu})^\star > 0$, which provide no information that is useful to distinguish $\underline{\nu}$ from problems $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$. A measure of this useful information is the Kullback-Leibler divergence between the laws of the rewards after T rounds, $\mathbb{P}_{\underline{\nu},T}$ and $\mathbb{P}_{\underline{\nu}',T}$ when the underlying problems are, respectively, $\underline{\nu}$ and $\underline{\nu}'$. This total Kullback-Leibler divergence is computed thanks to a chain rule, see Garivier et al. [2019],

$$\mathrm{KL}\left(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\right) = \sum_{t=1}^{T} \mathbb{E}_{\underline{\nu}}\big[\mathrm{KL}(\nu_{A_t}, \nu'_{A_t}))\big] = \sum_{t=1}^{T} \mathbb{E}_{\underline{\nu}}\left[\sum_{a=1}^{K} p_{t,a}\, \mathrm{KL}(\nu_a, \nu'_a)\right] .$$

This quantity can be factored as a sum over the availalble probabilities $\underline{p} \in \mathcal{P}$.

$$\sum_{\underline{p}\in\mathcal{P}} \mathbb{E}_{\underline{\nu}}\big[N^{(\underline{p})}(T)\big] \sum_{a\in[K]} p_a \operatorname{KL}(\nu_a, \nu'_a) = \ln T \bigg(\sum_{\underline{p}\in\mathcal{P}} n_T(\underline{p}) \sum_{a\in[K]} p_a \operatorname{KL}(\nu_a, \nu'_a) \bigg) .$$

When ν' belongs to $\textsc{Alt}(\underline{\nu})$, only the coordinates such that $\underline{p}^\star(a)(\underline{\nu}) = 0$ contribute to the sum over $a \in [K]$, and this can be further rewritten as

$$\ln T \bigg(\sum_{\underline{p}\in\mathcal{P}} n_T(\underline{p}) \sum_{\substack{a\in[K] \\ p^\star_a(\underline{\nu})=0}} p_a \operatorname{KL}(\nu_a, \nu'_a) \bigg) .$$

Asymptotically, the algorithm must maintain this amount of information above $\ln T$ in order to satisfy the UFC assumption, see Lemma 5.2 below. This puts a constraint on the limit of $n_t(\underline{p})$ for all $\underline{p}$, see (5.23), which translates into a lower bound on the regret thanks to (5.22).

Theorem 5.6. *For all algorithms that are UFC over $\mathcal{D}$, and for all problems $\underline{\nu} = (\nu_1, \dots, \nu_K)$ in $\mathcal{D}$ with a unique optimal action $\underline{p}^\star(\underline{\nu})$,*

$$\liminf_{T\to\infty} \frac{R_T(\underline{\nu})}{\ln T} \geqslant c(\mathcal{P}, \underline{\nu}) ,$$

where $c(\mathcal{P}, \underline{\nu}) \in [0, +\infty)$ is defined as the constrained infimum:

$$\inf_{n\in\mathbb{R}_+^{\mathcal{P}}} \sum_{\underline{p}\in\mathcal{P}} \Delta(\underline{p})\, n(\underline{p}) \quad \textit{under the constraint that}$$

$$\forall\, \underline{\nu}' \in \textsc{Alt}(\underline{\nu}), \qquad \sum_{\substack{\underline{p}\in\mathcal{P} \\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n(\underline{p}) \sum_{\substack{a\in[K] \\ p^\star_a(\underline{\nu})=0}} p_a \operatorname{KL}(\nu_a, \nu'_a) \geqslant 1 . \quad (5.23)$$

Proof. Let us denote $\underline{p}^\star = \underline{p}^\star(\underline{\nu})$, omitting the dependence on $\underline{\nu}$. We will lower bound all cluster points of the sequence $(R_t/\ln t)$. Consider an increasing sequence (t_m) such that $\big(R_{t_m}/\ln t_m\big)$ converges to a cluster point in $\mathbb{R}_+ \cup \{+\infty\}$. If the only cluster point is $+\infty$, the claim trivially holds. Otherwise, assume that the limit is finite.

Since for all $\underline{p} \neq \underline{p}^\star$, the quantities $\Delta(\underline{p})$ are positive and $n_{t_m}(\underline{p})$ are non-negative, all the sequences $\big(n_{t_m}(\underline{p})\big)_{m\in\mathbb{N}}$ for $\underline{p} \neq \underline{p}^\star$ are bounded. Therefore, we may extract a subsequence from (t_m), which we call (T_m), such that the sequences $\big(n_{T_m}(p)\big)_{m\in\mathbb{N}}$ converge for all $\underline{p} \neq \underline{p}^\star$, and we call $n(\underline{p})$ their limits. In the rest of the proof, we will carefully avoid refering to the undefined quantity $n(\underline{p}^\star)$.

Let $\underline{\nu}'$ in $\mathcal{D}$ be an alternative bandit problem, and denote by $\mathbb{P}_{\underline{\nu},t}$ (respectively $\mathbb{P}_{\underline{\nu}',t}$), the law of the rewards $Y_1, \dots, Y_t$, under problem $\underline{\nu}$ (respectively $\underline{\nu}'$). As explained in the paragraph preceding the proof,

$$\frac{\operatorname{KL}\left(\mathbb{P}_{\underline{\nu},t}, \mathbb{P}_{\underline{\nu}',t}\right)}{\ln t} = \sum_{\underline{p}\in\mathcal{P}} n_t(\underline{p}) \sum_{a=1}^{K} p_a \operatorname{KL}(\nu_a, \nu'_a) .$$

Then by Lemma 5.2 below, for any problem $\underline{\nu}'$ in which $\underline{p}^\star$ is suboptimal,

$$\liminf_{t\to\infty} \sum_{\underline{p}\in\mathcal{P}} n_t(\underline{p}) \sum_{a=1}^{K} p_a \operatorname{KL}(\nu_a, \nu'_a) \geqslant 1 . \quad (5.24)$$

Moreover, if $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$, we have either $p_a^\star = 0$ or $\nu_a = \nu'_a$, therefore

$$\sum_{a=1}^{K} p_a^\star \operatorname{KL}(\nu_a, \nu'_a) = 0\,,$$

and we may remove the term corresponding to $\underline{p} = \underline{p}^\star$ from the sum (5.24). Thus, considering the convergent sequences $\big(n_{T_m}(\underline{p})\big)$, and by identity of the lim inf, we deduce that for all $\underline{\nu}' \in \textsc{Alt}(\nu)$,

$$\sum_{\substack{\underline{p}\in\mathcal{P}\\ \underline{p}\neq\underline{p}^\star}} n(\underline{p}) \sum_{a=1}^{K} p_a \operatorname{KL}(\nu_a, \nu'_a) = \sum_{\substack{\underline{p}\in\mathcal{P}\\ \underline{p}\neq\underline{p}^\star}} n(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star=0}} p_a \operatorname{KL}(\nu_a, \nu'_a) \geqslant 1\,,$$

where the equality holds since we know that $\operatorname{KL}(\nu_a, \nu'_a) = 0$ whenever $p_a^\star > 0$. Therefore n satisfies the constraint (5.23). In conclusion, we have shown that all cluster points of $(R_t/\ln t)$ are lower bounded by

$$\sum_{\underline{p}\in\mathcal{P}} \Delta(\underline{p})\, n(\underline{p}) \quad \text{for some } n : \mathcal{P} \to \mathbb{R}_+ \text{ satifsfying the constraint (5.23).}$$

Note that the quantity $n(\underline{p}^\star)$ does not influence the value of the lower bound as $\Delta(\underline{p}^\star) = 0$. This concludes the proof. □

The following lemma summarizes the information-theoretic computations underlying this lower bound.

Lemma 5.2 (Asymptotic divergence). *Let $\underline{\nu}$ be a bandit problem in a model $\mathcal{D}$ with a unique optimal action $\underline{p}^\star(\underline{\nu})$. Let $\underline{\nu}'$ be a bandit problem in $\mathcal{D}$ for which $\underline{p}^\star(\underline{\nu})$ is suboptimal. Then, if the algorithm used to select the actions is UFC over $\mathcal{D}$,*

$$\liminf_{T\to\infty} \frac{\operatorname{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big)}{\ln T} \geqslant 1\,,$$

where $\mathbb{P}_{\underline{\nu},T}$ and $\mathbb{P}_{\underline{\nu}',T}$ denote the law of the rewards $Y_1, \dots, Y_T$ when the underlying problems are, respectively, $\underline{\nu}$ and $\underline{\nu}'$.

Proof. Denote $\underline{p}^\star = \underline{p}^\star(\underline{\nu})$, omitting the dependence on $\underline{\nu}$; denote also by $\operatorname{kl}(\cdot,\cdot)$ the Bernoulli Kullback-Leibler divergence. By the data-processing inequality for [0,1]–valued random variables (see Section 2.1 in [Garivier et al., 2019]), and using the standard inequality $\operatorname{kl}(p,q) \geqslant p\ln(1/q) - \ln 2$,

$$\operatorname{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big) \geqslant \operatorname{kl}\left(\mathbb{E}_{\underline{\nu}}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right], \mathbb{E}_{\underline{\nu}'}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right]\right) \geqslant \mathbb{E}_{\underline{\nu}}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right] \ln\left(\frac{T}{\mathbb{E}_{\underline{\nu}'}\Big[N^{(\underline{p}^\star)}(T)\Big]}\right) - \ln 2\,.$$

Since the strategy is UFC, and since $\underline{p}^\star$ is the only optimal action for $\underline{\nu}$,

$$\frac{R_T(\underline{\nu})}{T} \xrightarrow[T\to\infty]{} 0, \quad \text{thus} \quad \mathbb{E}_{\underline{\nu}}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right] \xrightarrow[T\to\infty]{} 1.$$

Next, for all $\alpha > 0$, there exists a constant $C > 0$, possibly depending on $\alpha, \underline{\nu}$ and $\underline{\nu}'$ such that for all times T, we have $R_T(\underline{\nu}') \leqslant CT^\alpha$. Therefore, as $\underline{p}^\star$ is suboptimal in $\underline{\nu}'$,

$$R_T(\underline{\nu}') \geqslant \Delta_{\underline{\nu}'}(\underline{p}^\star)\, \mathbb{E}_{\underline{\nu}'}\Big[N^{(\underline{p}^\star)}(T)\Big], \quad \text{so that} \quad \mathbb{E}_{\underline{\nu}'}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right] \leqslant \frac{C}{\Delta_{\underline{\nu}'}(\underline{p}^\star)} T^{\alpha-1} \stackrel{\text{def}}{=} C' T^{\alpha-1},$$

where we denoted by $\Delta_{\underline{\nu}'}(\underline{p}^\star)$ the gap of $\underline{p}^\star$ in $\underline{\nu}'$ and where we defined a modified constant C'. Combining the previous inequalities, we obtain

$$\frac{\mathrm{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big)}{\ln T} \geqslant \underbrace{\mathbb{E}_{\underline{\nu}}\left[\frac{N^{(\underline{p}^\star)}(T)}{T}\right]}_{\longrightarrow 1} \underbrace{\frac{\ln\left(T^{1-\alpha}/C'\right)}{\ln T}}_{\longrightarrow 1-\alpha} - \underbrace{\frac{\ln 2}{\ln T}}_{\longrightarrow 0}.$$

Therefore, as $T \to \infty$, we have

$$\liminf_{T\to\infty} \frac{\mathrm{KL}\big(\mathbb{P}_{\underline{\nu},T}, \mathbb{P}_{\underline{\nu}',T}\big)}{\ln T} \geqslant 1-\alpha\,.$$

The claimed result follows by taking $\alpha \to 0$. □

5.5.2. Discussion of the lower bound / Finite $\mathcal{P}$

This lower bounds shows in particular that the regret of a UFC algorithm on any bandit problem $\underline{\nu}$ such that $c(\mathcal{P}, \underline{\nu}) > 0$ grows at least logarithmically as $T \to \infty$. On the other hand, when $c(\mathcal{P}, \underline{\nu}) = 0$, this lower bound does not say anything. However, the optimality of this type of lower bound in other problems, e.g., in Combes et al. [2017], suggests that it should be possible to have sub-logarithmic regret in that case, i.e., to have the true limit

$$\frac{R_T(\underline{\nu})}{\ln T} \xrightarrow[T\to\infty]{} 0 \quad \text{whenever} \quad c(\mathcal{P}, \underline{\nu}) = 0\,.$$

In fact we already knew from Theorem 5.4 that bounded regret, a guarantee stronger that sub-logarithmic regret, is sometimes possible, e.g., when $\mathcal{P}$ is included in the interior of the simplex.

To develop this comparison, let us now state simple conditions to check whether $c(\mathcal{P}, \underline{\nu})$ is positive.

Proposition 5.3. *In the setting and under the conditions of Theorem 5.6, we have the equivalence*

$$c(\mathcal{P}, \underline{\nu}) = 0 \quad \textit{if and only if} \quad \textsc{Alt}(\underline{\nu}) \textit{ is empty}.$$

Note that if $p_a^\star > 0$ for all $a \in [K]$, then $\textsc{Alt}(\underline{\nu})$ is empty.

Proof. If $\textsc{Alt}(\underline{\mu})$ is empty then the linear program is unconstrained and $c(\mathcal{P}, \underline{\nu})$ is 0. For the converse statement, assume by contradiction that $\textsc{Alt}(\underline{\nu})$ is non-empty and that the infimum is 0. For $\varepsilon > 0$, there exists $n^\varepsilon \in \mathbb{R}_+^{\mathcal{P}}$, satisfying the constraint and such that

$$\sum_{\underline{p}\in\mathcal{P}} \Delta(\underline{p})\, n^\varepsilon(\underline{p}) \leqslant \varepsilon\,.$$

Then for any $\underline{\nu}' \in \textsc{Alt}(\underline{\nu})$, since n^ε satisfies the constraint (5.23),

$$\sum_{\substack{\underline{p}\in\mathcal{P}\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n^\varepsilon(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a, \nu'_a) \geqslant 1\,.$$

And therefore,

$$\begin{aligned} 1 \leqslant \sum_{\substack{\underline{p}\in\mathcal{P}\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} n^\varepsilon(\underline{p}) \sum_{\substack{a\in[K]\\ p_a^\star=0}} p_a \,\mathrm{KL}(\nu_a, \nu'_a) &\leqslant \max_{\underline{p}\in\mathcal{P},\, \underline{p}\neq\underline{p}^\star(\underline{\nu})} \left\{ \frac{1}{\Delta(\underline{p})} \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a, \nu'_a) \right\} \sum_{\substack{\underline{p}\in\mathcal{P}\\ \underline{p}\neq\underline{p}^\star(\underline{\nu})}} \Delta(\underline{p})\, n^\varepsilon(\underline{p}) \\ &\leqslant \varepsilon \max_{\underline{p}\in\mathcal{P},\, \underline{p}\neq\underline{p}^\star(\underline{\nu})} \left\{ \frac{1}{\Delta(\underline{p})} \sum_{\substack{a\in[K]\\ p_a^\star(\underline{\nu})=0}} p_a \,\mathrm{KL}(\nu_a, \nu'_a) \right\} \end{aligned}$$

This leads to a contradiction for ε small enough. □

Recall that we denote by I the range of means for distributions in the model. When I is not bounded from above we get a simpler characterization that depends only on the optimal action: the lower bound is null if and only $\underline{p}^\star(\underline{\nu})$ lies in the interior of the simplex.

Proposition 5.4. *In the setting and under the assumptions of Theorem 5.6, and under the additional assumption that the range of possible means I is not bounded from above, the following equivalences hold:*

$$c(\mathcal{P}, \underline{\nu}) = 0 \quad \Leftrightarrow \quad \textsc{Alt}(\underline{\nu}) \text{ is empty} \quad \Leftrightarrow \quad p^\star_a(\underline{\nu}) > 0 \quad \text{for all } a \in [K]\,.$$

Proof. The only implication left to prove is that if $p^\star_a(\underline{\nu}) = 0$ for some $a \in [K]$, then $\textsc{Alt}(\underline{\nu})$ is non-empty. We assumed with no loss of generality (see page 182) that for each arm k, there exists $\underline{p} \in \mathcal{P}$ with $p_k > 0$. There exists in particular $\underline{p} \in \mathcal{P}$ with $p_a > 0$. Since $\underline{p}^\star(\nu)$ is the unique optimal arm of ν, the gap $\Delta(p)$ is positive. Let $\underline{\nu}'$ be a bandit problem in $\mathcal{D}$, and $\alpha > \Delta(\underline{p})/p_a$ be a parameter such that

$$\mathrm{E}(\nu'_i) = \begin{cases} \mathrm{E}(\nu_i) & \text{if } i \neq a \\ \mathrm{E}(\nu_i) + \alpha & \text{if } i = a \end{cases}\,.$$

The existence of such a problem is guaranteed for some α large enough by the assumption that the means of $\mathcal{D}$ are not bounded by above.

It only remains to show that $\underline{p}^\star(\underline{\nu})$ is suboptimal for $\underline{\nu}'$, which is true as

$$\langle \underline{p}, \underline{\mu}' \rangle = \langle \underline{p}, \underline{\mu} \rangle + \alpha\, p_a > \langle \underline{p}, \underline{\mu} \rangle + \frac{\Delta(\underline{p})}{p_a}\, p_a = \langle \underline{p}^\star(\underline{\nu}), \underline{\mu} \rangle = \langle \underline{p}^\star(\underline{\nu}), \underline{\mu}' \rangle\,,$$

proving the claim. □

A natural example of a (tractable) model with unbounded means is the Gaussian noise case. More generally, one can consider reward distributions lying in a one-dimensional exponential family, which can sometimes have unbounded means; see, e.g., Cappé et al. [2013].

This last characterization is another argument in favor of our conjecture in Open question 5.3: modulo some small technical conditions, bounded regret should be possible whenever the optimal action lies in the interior of the simplex. More generally, it would be nice to have an algorithm matching the asymptotic lower bound presented here, with the ultimate goal of designing an algorithm enjoying both asymptotically optimal and reasonable finite-time time guarantees. To this end, approaches developed in Lattimore and Szepesvári [2017], Combes et al. [2017], Hao et al. [2020] seem promising; there are however significant technical obstacles to overcome in order to adapt these approaches.

5.5.3. Extension to polytopes $\mathcal{P}$ (by a reduction argument)

We now show how to extend the lower bound from the case of a finite $\mathcal{P}$ to the case of a (closed) polytope $\mathcal{P}$, i.e., by the Krein-Milman theorem, the convex hull of a finite set of extremal points $\mathrm{Ext}(\mathcal{P})$. To do so, we show that any strategy playing in $\mathcal{P}$ can be transformed into a strategy suffering the same regret, but playing exclusively in $\mathrm{Ext}(\mathcal{P})$.

We know, via Krein-Milman's theorem, that $\mathcal{P}$ is the convex hull of $\mathrm{Ext}(\mathcal{P})$ (remember we assumed that $\mathcal{P}$ is closed). Therefore, by Carathéodory's theorem, any point in $\mathcal{P}$ can be written as a finite convex combination of points in $\mathrm{Ext}(\mathcal{P})$. We can thus build a mapping Φ from $\mathcal{P}$ to the set of probability measures on $\mathrm{Ext}(\mathcal{P})$ such that the expected value of $\Phi(\underline{p})$ is $\underline{p}$. Note that the proof of Carathéodory's theorem provides a concrete way to build this mapping.

More precisely, from the mentioned convex-hull property, any point in $\mathcal{P}$ can be written as a finite convex combination of points in $\mathrm{Ext}(\mathcal{P})$. We consider a convex decomposition with elements in $\mathrm{Ext}(\mathcal{P})$ as a probability measure on $\mathrm{Ext}(\mathcal{P})$. The proof of Caratheodory's theorem provides a concrete way of constructing the needed convex decompositions / probability distributions. This leads to a mapping Φ from $\mathcal{P}$ to the set of probability measures on $\mathrm{Ext}(\mathcal{P})$ such that the expected value of $\Phi(\underline{p})$ is $\underline{p}$.

For a given strategy ψ, let $\Phi \cdot \psi$ be the strategy defined as follows: at time t, given a history of plays and observations $\underline{p}_1, A_1, Y_1 \ldots, \underline{p}_{t-1}, A_{t-1}, Y_{t-1}$, if ψ picks $\underline{p}_t \in \mathcal{P}$ at time t, then $\Phi \cdot \psi$ samples $\underline{q}_t \in \mathrm{Ext}(\mathcal{P})$ with law $\underline{q}_t \sim \Phi(\underline{p}_t)$. This mapping preserves the laws of the chosen pure actions, and thus of the rewards. Hence the expected regret is preserved:

$$R_T(\psi, \underline{\nu}) = R_T\big(\Phi \cdot \psi, \underline{\nu}\big)\,.$$

In particular, if ψ is UFC over a bandit model $\mathcal{D}$, then $\Phi \cdot \psi$ is also UFC over $\mathcal{D}$. Since $\Phi \cdot \psi$ plays only in the finite set $\mathrm{Ext}(\mathcal{P})$, we can apply Theorem 5.6 and deduce that

$$\liminf_{T\to\infty} R_T(\psi, \underline{\nu}) \geqslant c\big(\mathrm{Ext}(\mathcal{P}), \underline{\nu}\big)\,,$$

where c is the value of the optimization problem defined in Theorem 5.6.

5.6. Some numerical experiments on synthetic data

Setup, probability set and problems considered. In this section, we perform some numerical experiments that support our conjecture that UCB enjoys bounded regret whenever the optimal probability lies in the interior of the simplex (Open question 5.3). To this end, we consider a simple probability set in dimension $K = 3$ that is not completely included in the interior of the simplex, namely, the set, for $\ell \in (0, 1/2)$,

$$\mathcal{P} = \big\{(p_1, p_2, p_3) \in \mathcal{S} \;:\quad p_1 \geqslant \ell \quad \text{and} \quad p_2 \geqslant \ell\big\}\,.$$

This action set has 3 corners

$$\underline{p}_1 = \big(1-\ell, \ell, 0\big), \quad \underline{p}_2 = \big(\ell, 1-\ell, 0\big), \text{ and } \quad \underline{p}_3 = \big(\ell, \ell, 1-2\ell\big)\,,$$

and out of these three corners, only $\underline{p}_3$ lies in the interior of the simplex. We consider a family of bandit problems $\underline{\nu}_\alpha$, with $\alpha \in (-1/6, 1/6)$, and Bernoulli distributions, defined by their means:

$$\underline{\mu}_\alpha = (1/2 + \alpha, 1/3, 1/2 - \alpha)\,.$$

These problems are defined so that if $\alpha < 0$, then $\underline{p}_3$ is the only optimal action, whereas if $\alpha > 0$, then $\underline{p}_1$ is the only optimal action. Indeed, we have

$$\begin{aligned}
\big\langle \underline{p}_3 - \underline{p}_1, \underline{\mu}_\alpha \big\rangle &= \big(2\ell - 1\big)\Big(\frac{1}{2} + \alpha\Big) + 0 + \big(1 - 2\ell\big)\Big(\frac{1}{2} - \alpha\Big) = -2\alpha(1-2\ell) \\
\big\langle \underline{p}_1 - \underline{p}_2, \underline{\mu}_\alpha \big\rangle &= (1-2\ell)\Big(\frac{1}{2} + \alpha\Big) + (2\ell - 1)\frac{1}{3} + 0 = (1-2\ell)\Big(\frac{1}{6} + \alpha\Big) > 0 \\
\big\langle \underline{p}_3 - \underline{p}_2, \underline{\mu}_\alpha \big\rangle &= 0 + (2\ell - 1)\frac{1}{3} + (1-2\ell)\Big(\frac{1}{2} - \alpha\Big) = (1-2\ell)\Big(\frac{1}{6} - \alpha\Big) > 0.
\end{aligned}$$

By Theorem 5.6, Proposition 5.4 and the reduction argument of Section 5.5.3, any UFC algorithm, and in particular UCB, must suffer logarithmic regret on the problems $\underline{\nu}_\alpha$ when $\alpha > 0$. On the

other hand, the lower bound is void for $\underline{\nu}_\alpha$ when $\alpha < 0$. In that case, we conjectured that UCB should obtain bounded regret. To sum up (see also Figure 5.2)

$$\begin{aligned} \alpha < 0 \quad &\Rightarrow \quad \underline{p}_3, \text{ in the interior of the simplex, is the only optimal action} \\ &\Rightarrow \quad \text{no lower bound, and bounded regret seems possible}, \end{aligned}$$

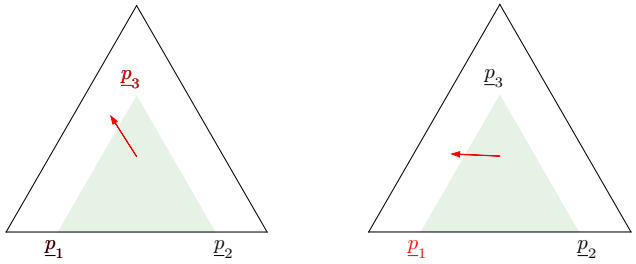
whereas

$$\begin{aligned} \alpha > 0 \quad &\Rightarrow \quad \underline{p}_1, \text{ on the border of the simplex, is the only optimal action} \\ &\Rightarrow \quad \text{logarithmic lower bound on the regret} \ . \end{aligned}$$

In our experiments, we set $\ell = 0.1$ and let α vary in $\{-0.1, -0.05, 0.05, 0.1\}$. We run the diversity-preserving UCB algorithm on all problems $\underline{\nu}_\alpha$, over $T = 20{,}000$ time steps for $N = 75$ runs. The regret suffered by the algorithm is estimated with

$$\widehat{R}_T(\underline{\nu}_\alpha) = \frac{1}{N}\sum_{i=1}^{N} \widehat{R}_T(\underline{\nu}_\alpha, i)\,, \quad \text{where} \quad \widehat{R}_T(\underline{\nu}_\alpha, i) = \sum_{t=1}^{T} \big\langle \underline{p}^\star(\underline{\nu}_\alpha) - \underline{p}_t(\alpha, i),\, \underline{\mu}_\alpha \big\rangle$$

and where we denoted by $\underline{p}_t(\alpha, i)$ the mixed action at round t on problem $\underline{\nu}_\alpha$ on the i–th run.



(a) $\underline{p}^\star = \underline{p}_3$ is in the interior

(b) $\underline{p}^\star = \underline{p}_1$ is on the border

Figure 5.2.: The probability set $\mathcal{P}$ and two different bandit problems. On the left, the optimal action is $\underline{p}_3$, which is in the interior of the simplex. On the right, the optimal action (in red) is $\underline{p}_1$, which is on the border on the simplex. We also plot in red the projection of the mean-payoff vector (anchored at the origin) on the simplex.

Figure 5.3 reports the results of our experiments. As expected, the algorithm yields logarithmic regret for $\underline{\nu}_\alpha$ if and only if $\alpha < 0$, i.e., whenever the optimal probability is on the border simplex.

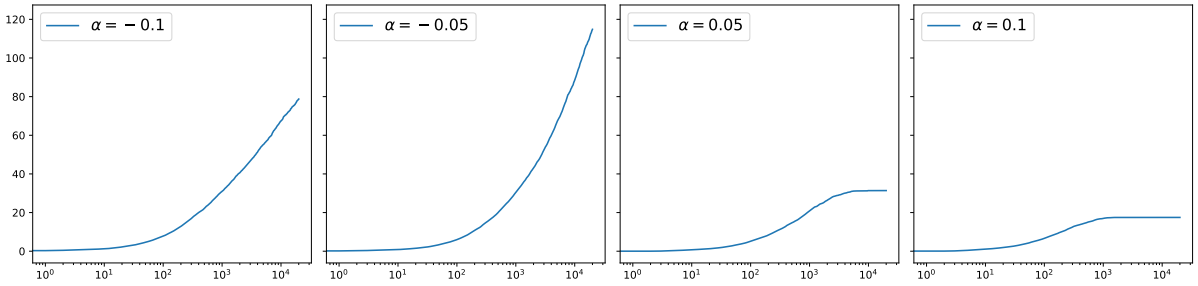


Figure 5.3.: Performance of diversity-preserving UCB on different problems $\underline{\nu}_\alpha$, with the parameter $\alpha \in \{-0.1, -0.05, 0.05, 0.1\}$. Each algorithm was run $N = 75$ times, each run lasting $T = 20{,}000$ time steps. Solid lines report the values of the estimated regrets, while shaded areas correspond to ± 2 standard errors of the estimates.

Bibliography

Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS '11)*, pages 2312–2320, 2011.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995a.

Rajeev Agrawal. Sample mean based index policies with $O(logn)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995b.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems (NeurIPS '19)*, pages 9252–9262, 2019.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 37(4):1591–1646, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory (COLT'09)*, pages 217–226, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Conference on Learning Theory (COLT'07)*, pages 454–468, 2007.

Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for Thompson sampling. In *Advances in Neural Information Processing Systems (NIPS'13)*, pages 638–646, 2013.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011a.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011b.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the Lipschitz constant. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT'11)*, pages 144–158. Springer, 2011c.

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT'18)*, volume 83 of PMLR, pages 111–127, 2018.

Adam D. Bull. Adaptive-treed bandits. *Bernoulli*, 21(4):2289–2307, 2015.

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

T. Tony Cai. Minimax and adaptive inference in nonparametric function estimation. *Statistical Science*, 27(1):31–50, 2012.

T. Tony Cai and Mark G. Low. On adaptive estimation of linear functionals. 33(5):2311–2343, 2005.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI '12)*, UAI'12, pages 142–151, 2012.

L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 160–169. Association for Computing Machinery, 2019.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi and Ohad Shamir. Bandit regret scaling with the effective loss range. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT'18)*, volume 83 of PMLR, pages 128–151, 2018.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Yuan Shih Chow and Henry Teicher. *Probability Theory*. Springer, 1988.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AIStats'11)*, pages 208–214, 2011.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS'17)*, pages 1763–1771, 2017.

Pierre-Arnaud Coquelin and Rémi Munos. Bandit algorithms for tree search. In *Proceedings of the 23th Conference on Uncertainty in Artificial Intelligence (UAI '07)*, 2007.

Wesley Cowan and Michael N. Katehakis. An asymptotically optimal policy for uniform bandits of unknown support. arXiv 1505.01918, 2015.

Wesley Cowan, Junya Honda, and Michael N. Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

Steven De Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Rémy Degenne and Vianney 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.

Rémy Degenne, Evrard Garcelon, and Vianney Perchet. Bandits with side observations: Bounded vs. logarithmic regret. arXiv 1807.03558, 2018.

Joseph L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th ACM Symposium on Theory of Computing (STOC'97)*, pages 334–343, 1997.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory (COLT'11)*, pages 359–376, 2011.

Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. arXiv 1805.05071, 2018.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems (NIPS'16)*, pages 1198–1206, 2016.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems (NIPS'18)*, pages 2600–2609, 2018.

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

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems (NIPS'15)*, pages 667–675, 2015.

Hédi Hadiji. Polynomial cost of adaptation for $\mathcal{X}$-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS'19)*, pages 1029–1038, 2019.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in K-armed bandits. arXiv 2006.03378, 2020.

Hédi Hadiji, Sébastien Gerchinovitz, Jean-Michel Loubes, and Gilles Stoltz. Diversity-preserving k-armed bandits, revisited. arxiv 2010.01874, 2020.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AIStats'20)*, volume 108, pages 3536–3545, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85:361–391, 2011.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16(113):3721–3756, 2015.

Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the Lambert W function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):Article 51, 2008.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems (NIPS'16)*, pages 325–333, 2016.

Olav Kallenberg. *Foundations of Modern Probability*. Springer Science & Business Media, 2006.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats '12)*, AIStats'12, pages 592–600, 2012.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. arXiv 2002.11182, 2020.

Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.

Robert Kleinberg. Nearly Tight Bounds for the Continuum-armed Bandit Problem. In *Proceedings of the 17th International Conference on Neural Information Processing Systems (NIPS'14)*, pages 697–704. MIT Press, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems (NIPS'14)*, pages 613–621, 2014.

Wouter M. Koolen. Adaftrl. Blog post, Oct 2016. URL `http://blog.wouterkoolen.info/AdaFTRL/post.html`.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1–dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS'13)*, pages 1448–1456, 2013.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Proceedings of the 32nd Conference on Learning Theory (COLT'19)*, volume 99 of PMLR, pages 2025–2027, 2019.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

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

Tor Lattimore. Regret analysis of the anytime optimally confident ucb algorithm. arXiv 1603.08661, 2016.

Tor Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Advances in Neural Information Processing Systems (NIPS'17)*, pages 1584–1593, 2017.

Tor Lattimore and Csaba Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AIStats'20)*, pages 728–737, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Oleg Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Oleg V. Lepskii. Estimation of the maximum of a nonparametric signal to within a constant. *Theory of Probability & Its Applications*, 38(1):152–158, 1994. doi: 10.1137/1138013.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 2019.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 661–670, 2010.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. Calibrated fairness in bandits. In *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (Fat/ML 2017)*, 2017.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in $\mathcal{X}$–armed bandits. In *Proceedings of the 31st Conference On Learning Theory (COLT'18)*, volume 75 of PMLR, pages 1463–1492, 2018.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th Conference on Learning Theory (COLT'11)*, 2011.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume XXXIII of *Ecole d'Eté de Probabilités de Saint-Flour*. Springer, 2007. Lectures given in 2003, published in 2007.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT '17)*, 2017.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. arXiv 1907.10516, 2019.

Mark D. Reid, Rafael M. Frongillo, Robert C. Williamson, and Nishant Mehta. Generalized mixability via entropic duality. In *Proceedings of the 28th Conference on Learning Theory (COLT'15)*, volume 40 of *Proceedings of Machine Learning Research*, pages 1501–1522. PMLR, 2015.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 30th Conference on Learning Theory (COLT'17)*, volume 65 of PMLR, pages 1743–1759, 2017.

Xuedong Shang, Emilie Kaufmann, and Michal Valko. General parallel optimization without a metric. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory (ALT'19)*, volume 98, 2019.

Gordon Simons, Lijian Yang, and Yi-Ching Yao. Doob, Ignatov and optional skipping. *Annals of Probability*, 30(4):1933–1958, 2002.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic Simultaneous Optimistic Optimization. In *International Conference on Machine Learning*, 2013.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory (COLT'18)*, volume 75 of PMLR, pages 1263–1291, 2018.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AIStats'20)*, volume 89 of PMLR, pages 467–475, 2019.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Sur quelques questions d'adaptation dans des problèmes de bandits stochastiques

Mots clés: Bandits stochastiques à plusieurs bras, algorithme Upper Confidence Bound (UCB), optimalité minimax, optimalité asymptotique, bandits à continuum de bras, statistiques adaptatives

Résumé: Cette thèse s'inscrit dans le domaine des statistiques séquentielles. Le cadre principal étudié est celui des bandits stochastiques à plusieurs bras, cadre idéal qui modélise le dilemne exploration-exploitation face à des choix répétés. La thèse est composée de quatre chapitres, précédés d'une introduction. Dans la première partie du corps de la thèse, on présente un nouvel algorithme capable d'atteindre des garanties optimales à la fois d'un point de vue distribution-dependent et distribution-free. Les deux chapitres suivants sont consacrés à des questions dites d'adaptation. D'abord, on propose un algorithme capable de s'adapter à la régularité inconnue dans des problèmes de bandits continus, mettant en évidence le coût polynomial de l'adaptation en bandits continus. Ensuite, on considère un problème d'adaptation au supports pour des problèmes de bandits à K bras, à distributions de paiements bornés dans des intervalles inconnus. Enfin, dans un dernier chapitre un peu à part, on étudie un cadre légèrement différent de bandits préservant la diversité. On montre que le regret optimal dans ce cadre croît à des vitesses différentes des vitesses classiques, avec notamment la possibilité d'atteindre un regret constant sous certaines hypothèses.

Title: On some adaptivity questions in stochastic multi-armed bandits

Keywords: Stochastic multi-armed bandits, upper confidence bound (UCB), minimax optimality, asymptotic optimality, continuum-armed bandits, adaptive statistics

Abstract: The main topics adressed in this thesis lie in the general domain of sequential learning, and in particular stochastic multi-armed bandits The thesis is divided into four chapters and an introduction. In the first part of the main body of the thesis, we design a new algorithm achieving, simultaneously, distribution-dependent and distribution-free optimal guarantees. The next two chapters are devoted to adaptivity questions. First, in the context of continuum-armed bandits, we present a new algorithm which, for the first time, does not require the knowledge of the regularity of the bandit problem it is facing. Then, we study the issue of adapting to the unknown support of the payoffs in bounded K-armed bandits. We provide a procedure that (almost) obtains the same guarantees as if it was given the support in advance. In the final chapter, we study a slightly different bandit setting, designed to enforce diversity-preserving conditions on the strategies. We show that the optimal regert in this setting at a speed that is quite different from the traditional bandit setting. In particular, we observe that bounded regret is possible under some specific hypotheses.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France