Bandits Games and Clustering Foundations

Sébastien Bubeck

UNIVERSITÉ LILLE 1 — UFR SCIENTIFIQUE DE LILLE

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ LILLE 1

Spécialité : **Mathématiques**

présentée par

Sébastien BUBECK

JEUX DE BANDITS ET FONDATIONS DU CLUSTERING

Rapporteurs :	M. Olivier	**CATONI**	CNRS et ENS
	M. Nicolò	**CESA-BIANCHI**	Università degli Studi di Milano

Soutenue publiquement le **10 juin 2010** devant le jury composé de

Mme. Cristina	**BUTUCEA**	Université Lille 1	Co-Directrice
M. Olivier	**CATONI**	CNRS et ENS	Rapporteur
M. Nicolò	**CESA-BIANCHI**	Università degli Studi di Milano	Rapporteur
M. Pascal	**MASSART**	Université Paris-Sud	Examinateur
M. Eric	**MOULINES**	Télécom ParisTech	Examinateur
M. Rémi	**MUNOS**	INRIA Lille	Directeur

Acknowledgements – Remerciements

Je tiens avant tout à te remercier Rémi. Travailler avec toi fut un réel plaisir, ton enthousiasme permanent, ta disponibilité et ta vision mathématique unique me resteront longtemps en mémoire.

Jean-Yves, nous avons tout juste commencé à explorer nos centres d'intérêt communs, et j'ai le sentiment qu'il nous reste encore beaucoup de choses à faire. Je tenais particulièrement à te remercier de partager tes idées toujours très stimulantes (pour ne pas dire plus ...) avec moi.

Les us et coutumes du monde académique peuvent parfois être difficile à pénétrer, heureusement dans ce domaine j'ai eu comme maitre un expert en la matière, Gilles. Au niveau mathématique tu m'as permis de débuter ma thèse sur des bases solides, et ton aide a été inestimable.

I was lucky enough to be introduced to the world of research by you Ulrike. You taught me how to do (hopefully) useful theoretical research, but also all the basic tricks that a researcher has to know. I wish both of us had more time to continue our exciting projects, but I am confident that in the near future we will collaborate again!

In the cold and icy land of Alberta, lived a man known for his perfect knowledge of the right references, but also for his constant kindness. Csaba, I am looking forward to (finally) start a new project with you.

Gábor, we just started working together, and I hope that our collaboration will be fruitful. I wanted to thank you for welcoming me in Barcelona next year, according to my short visit I think we will have fun.

I also wanted to thank the Sequel team, in unalphabetical order: Daniil, Mohammad, Manuel, Alessandro, Phillipe, Jérémy, Victor, Alexandra, Odalric and Pierre-Arnaud. I enjoyed your challenging questions, or inspiring answers to my questions. I hope to continue this game with you, even if we are not next door anymore!

Special thanks to my referees, Nicolò and Olivier.

Bien que cela soit difficile, je tenais aussi à faire des remerciements à ma famille et mes amis d'enfance (ou d'adolescence). Mes pensées vont en premier aux p'tits, Arnaud, Thomas, Mathias, avec qui il est difficile de ne pas s'amuser, et bien sûr à Martin et Antoine, sans oublier Zacharie, Aude et Antoine, Louis, Anne-Catherine, Dave et Maggie (I hope you won't mind being in the french part of the acknowledgments!). Ma gratitude va à mes parents, pour avoir été un support constant dans tous les domaines.

Cette section ne serait pas complète sans remercier Phuket pour son inépuisable imagination pour les bêtises. Enfin, c'est à toi Anneso que je veux dédier ce travail.

Foreword

This thesis is divided into two parts. Chapter 2 to 7 concern bandits games and their extensions while Chapter 8 and 9 discuss new results on clustering. Each chapter can be read independently of others, introducing inevitably some redundancy. The organization is as follows:

- Chapter 1 is a short presentation (in French) of the contributions of this thesis.
- Chapter 2 is an introduction to the bandit theory. It presents the basic theoretical results with (for some of them) improved constants and/or simpler proofs. In this chapter we also review extensions of the basic multi-armed bandit game and highlight the contributions of this thesis to some of them.
- Chapter 3 presents two new algorithms for the multi-armed bandit game, MOSS (Minimax Optimal Strategy for the Stochastic case) and INF (Implicitly Normalized Forecaster). We prove that both algorithms are minimax optimal (up to a constant factor), solving a decade-long open problem. We also apply INF to other games (full information, label efficient, bandit label efficient, tracking the best expert), and improve the minimax rates for the regret whenever it is possible.
- Chapter 4 presents a new algorithm for the $\mathcal{X}$-armed bandit problem, HOO (Hierarchical Optimistic Optimization). We prove that, for metric spaces $\mathcal{X}$ with a well defined (sort of) metric dimension, HOO is minimax optimal (up to a logarithmic factor). We also prove, under weaker assumptions than any previous works, that it is possible to attain a regret of order $\sqrt{n}$ no matter the ambient dimension.
- Chapter 5 deals with the extension of bandits to the problem of planning in discounted and stochastic environments. We present a new algorithm, OLOP (Open Loop Optimistic Planning), and prove its minimax optimality (up to a logarithmic factor). We also show that OLOP attains much faster rate whenever the number of optimal sequences of actions is small.
- Chapter 6 introduces a new type of regret: the simple regret. In this chapter we study the links between the cumulative and simple regret and analyze the minimax rate for the simple regret.
- Chapter 7 further the study of the simple regret and proposes two new algorithms, SR (Successive Rejects) and UCB-E (Upper Confidence Bound Exploration). We prove that both algorithms have an optimal distribution-dependent rate of convergence to 0 for the simple regret up to a logarithmic factor.
- Chapter 8 presents a statistical view on clustering and introduces a new algorithm, NNC (Nearest Neighbor Clustering), which is the first provable algorithm to be asymptotically consistent for (almost) any objective function.
- Chapter 9 studies the ability of stability methods to select the number of clusters. In particular we consider the k-means algorithm and propose a new analysis of a non-trivial initialization scheme.

Contents

CHAPITRE 1

Introduction

Ce travail de thèse s'inscrit dans le domaine du *machine learning* et concerne plus particulièrement les sous-catégories de l'optimisation stochastique, du *online learning* et du *clustering*. Ces sous-domaines existent depuis plusieurs décennies mais ils ont tous reçu un éclairage différent au cours de ces dernières années. Notamment, les jeux de bandits offrent aujourd'hui un cadre commun pour l'optimisation stochastique et l'*online learning*. Ce point de vue conduit à de nombreuses extensions du jeu de base. C'est sur l'étude mathématique de ces jeux que se concentre la première partie de cette thèse. La seconde partie est quant à elle dédiée au *clustering* et plus particulièrement à deux notions importantes : la consistance asymptotique des algorithmes et la stabilité comme méthode de sélection de modèles.

Ce premier chapitre est l'occasion de présenter brièvement le contexte des deux parties qui constituent le corps de la thèse ainsi que de résumer les contributions des différents chapitres.

Contents

1. Les jeux de bandits

Le jeu du bandit a une longue histoire, il a été introduit pour la première fois par Robbins [1952] dans un contexte stochastique puis par Auer et al. [2003] comme un jeu contre un adversaire malicieux. Les deux versions sont décrites en détails dans le Chapitre 2 et on les rappelle brièvement ici dans les figures 1 et 2. Le Chapitre 2 est aussi l'occasion de rappeler les résultats de base et d'en donner une version légèrement améliorée dans la majorité des cas (soit au niveau des constantes soit au niveau de la preuve).

Comme nous allons le voir tout au long de cette thèse, le jeu du bandit et ses variantes modélisent de nombreux problèmes concrets en mathématiques appliquées. On peut citer par exemple le placement de bandeaux publicitaires sur une page internet, la construction d'une intelligence artificielle pour le jeu de Go ou encore la recherche efficace d'une fréquence de communication pour un dialogue entre téléphones mobiles. Dans la suite nous mettrons l'accent sur l'analyse mathématique rigoureuse des différents jeux mais sans perdre de vue les applications concrètes. Ainsi les problèmes réels donnent naissance à des questions mathématiques dont les

Paramètres : nombre de bras (ou actions) K ; nombre de tours n avec $n \geqslant K \geqslant 2$.

A chaque tour $t = 1, 2, \ldots, n$;
(1) le joueur choisit $I_t \in \{1, \ldots, K\}$, éventuellement à l'aide d'une randomisation externe ;
(2) simultanément l'adversaire choisit un vecteur de gains $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$;
(3) le joueur reçoit la récompense $g_{I_t,t}$ (sans voir la récompense des autres bras).

But : Maximiser les gains cumulés. De manière (presque) équivalente on cherche à minimiser le regret (cumulé) $R_n = \max_{i \in \{1,\ldots,K\}} \sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t}$. Un objectif moins ambitieux est de minimiser le pseudo-regret (cumulé) $\overline{R}_n = \max_{i \in \{1,\ldots,K\}} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}$.

FIG. 1: Jeu du bandit.

Paramètres connus : nombre de bras K ; nombre de tours n avec $n \geqslant K \geqslant 2$;

Paramètres inconnus : K distributions de probabilités $\nu_1, \ldots, \nu_K$ sur $[0,1]$ (ayant pour moyennes $\mu_1, \ldots, \mu_K$).

A chaque tour $t = 1, 2, \ldots, n$;
(1) le joueur choisit $I_t \in \{1, \ldots, K\}$;
(2) l'environnement tire la récompense $Y_t \sim \nu_{I_t}$ indépendamment du passé étant donné I_t.

But : Maximiser en moyenne les gains cumulés. De manière équivalente on cherche à minimiser le pseudo-regret (cumulé) $\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \max_{i \in \{1,\ldots,K\}} \mu_i - \mu_{I_t}$.

FIG. 2: Jeu du bandit stochastique.

réponses (le plus souvent sous la forme d'algorithmes) doivent en retour aider à la résolution du problème intial. Cette association à double sens entre mathématiques et problématiques réelles est au cœur de cette thèse et a guidé le choix de la majorité des sujets considérés.

1.1. Vitesses minimax du jeu du bandit et de ses variantes. Une problématique classique en statistiques est d'obtenir des vitesses minimax exactes pour différents types de regrets. Concernant le jeu du bandit on cherche à caractériser la quantité

$$\inf \sup \overline{R}_n$$

où $\inf$ représente l'infimum par rapport aux stratégies du joueur et $\sup$ le supremum par rapport aux stratégies de l'adversaire (ou par rapport au choix des probabilités dans le cas du jeu stochastique). On sait depuis Auer et al. [1995] que cette quantité est plus petite que $\sqrt{2nK\log(K)}$ et plus grande que $\frac{1}{20}\sqrt{nK}$ (pour les deux types de jeu). Ainsi la vitesse minimax n'était connue qu'à un facteur logarithmique près, y compris dans le cas du jeu stochastique.

Dans le Chapitre 3 on propose deux nouveaux algorithmes, MOSS (*Minimax Optimal Strategy for the Stochastic case*) pour le cas stochastique et INF (*Implicitly Normalized Forecaster*) pour le jeu général, chacun d'eux atteignant un pseudo-regret de l'ordre de $\sqrt{nK}$ (à une constante près),

Jeux de prédiction :

Paramètres : nombre de bras (ou actions) K ; nombre de tours n avec $n \geqslant K \geqslant 2$.

A chaque tour $t = 1, 2, \ldots, n$;
(1) le joueur choisit $I_t \in \{1, \ldots, K\}$, éventuellement à l'aide d'une randomisation externe ;
(2) simultanément l'adversaire choisit un vecteur de gains $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$;
(3) Le joueur reçoit le gain $g_{I_t,t}$ (sans forcement l'observer). Il observe
- le vecteur de gains $(g_{1,t}, \ldots, g_{K,t})$ dans le jeu à **information complète**,
- le vecteur de gains $(g_{1,t}, \ldots, g_{K,t})$ si il le demande en sachant qu'il n'est pas autorisé à le demander plus de m fois pour un entier $1 \leqslant m \leqslant n$ fixé. C'est le jeu du ***label efficient***,
- seulement $g_{I_t,t}$ dans le jeu du **bandit**,
- seulement $g_{I_t,t}$ si il le demande en sachant qu'il n'est pas autorisé à le demander plus de m fois pour un entier $1 \leqslant m \leqslant n$ fixé. C'est le jeu du **bandit *label efficient***,

But : Minimiser le regret (cumulé) $R_n = \max_{i \in \{1,\ldots,K\}} \sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t}$ ou le pseudo-regret (cumulé) $\overline{R}_n = \max_{i \in \{1,\ldots,K\}} \mathbb{E} \sum_{t=1}^{n} g_{i,t} - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}$.

FIG. 3: Quatre jeux de prédiction.

donnant ainsi la réponse à un problème ouvert depuis plus d'une décennie.

Le Chapitre 3 est aussi l'occasion de discuter différentes variantes du jeu du bandit étudiées dans Cesa-Bianchi et al. [1997], Auer [2002], Cesa-Bianchi et al. [2005] et Allenberg et al. [2006]. On les rappelle ici brièvement avec la Figure 3. Pour l'ensemble de ces nouveaux jeux (à l'exception du jeu à information complète) nous améliorons les vitesses minimax connues grâce à l'algorithme INF et à un nouveau type d'analyse unifiée des différents jeux. Les résultats obtenus sont résumés dans la table 1 (où le jeu du bandit indifférent correspond à un jeu de bandit classique où l'adversaire doit choisir sa suite de vecteurs de gains avant que le jeu ne commence). Dans ce même chapitre on considère aussi un regret plus général où on ne se compare pas au meilleur bras mais à la meilleure stratégie qui change S fois de bras, où $0 \leqslant S \leqslant n$ est un entier fixé. Pour ce regret aussi on améliore les vitesses minimax par rapport à l'état de l'art.

1.2. Le jeu du bandit stochastique avec une infinité de bras. Une limitation majeure des jeux décrits dans la section précédente est l'hypothèse, *a priori* bénigne, $n \geqslant K$. Autrement dit on suppose qu'il y a au moins autant de tours de jeu que de bras, et même significativement plus pour que les résultats deviennent intéressants. Cette hypothèse est évidemment indispensable quand on ne possède aucune connaissance *a priori* sur le comportement des bras puisqu'il faut au moins pouvoir tester chacun d'eux. Cependant dans de nombreux problèmes pratiques il existe une certaine structure sur les bras, une information sur l'un peut donner des informations sur les bras "proches". Dans le Chapitre 4 on montre que sous des hypothèses convenable il est en fait possible de gérer une infinité de bras, i.e., d'avoir un pseudo-regret sous linéaire (en le nombre de tours) sans tester chaque bras. Plus précisément on s'intéresse au jeu du bandit stochastique avec un espace mesurable de bras $\mathcal{X}$ et tel que l'ensemble des probabilités sur les bras possède une structure interne qui contraint la forme de la fonction moyenne (la fonction qui associe à

	$\inf\sup \overline{R}_n$		$\inf\sup \mathbb{E}R_n$		R_n
	Borne inf.	Borne sup.	Borne inf.	Borne inf.	Borne sup.
Information Complète	$\sqrt{n\log K}$	$\sqrt{n\log K}$	$\sqrt{n\log K}$	$\sqrt{n\log K}$	$\sqrt{n\log(K\delta^{-1})}$
Label Efficient	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$\mathbf{n}\sqrt{\frac{\log \mathbf{K}}{\mathbf{m}}}$	$\mathbf{n}\sqrt{\frac{\log(\mathbf{K}\delta^{-1})}{\mathbf{m}}}$
Bandit Indifférent	$\sqrt{nK}$	$\sqrt{\mathbf{nK}}$	$\sqrt{nK}$	$\sqrt{\mathbf{nK}}$	$\sqrt{\mathbf{nK}\log(\delta^{-1})}$
Bandit	$\sqrt{nK}$	$\sqrt{\mathbf{nK}}$	$\sqrt{nK}$	$\sqrt{\mathbf{nK}\log \mathbf{K}}$	$\sqrt{\frac{\mathbf{nK}}{\log \mathbf{K}}\log(\mathbf{K}\delta^{-1})}$
Bandit *Label Efficient*		$\mathbf{n}\sqrt{\frac{\mathbf{K}}{\mathbf{m}}}$		$\mathbf{n}\sqrt{\frac{\mathbf{K}\log \mathbf{K}}{\mathbf{m}}}$	$\mathbf{n}\sqrt{\frac{\mathbf{K}}{\mathbf{m}\log \mathbf{K}}\log(\mathbf{K}\delta^{-1})}$

TAB. 1: En rouge les résultat où nous améliorons par rapport à l'état de l'art. Ces bornes sont toutes à une constante près. Les bornes de la dernière colonne sont vraies avec probabilité au moins $1-\delta$.

chaque bras sa récompense moyenne). Ces conditions s'appliquent par exemple au cas d'un espace métrique compact $\mathcal{X}$ et aux fonctions moyennes 1-Lipschitz par rapport à cette métrique. Ce travail fait suite à une série de papiers sur les bandits stochastiques avec infinité de bras. En particulier on généralise les résultats de Kleinberg [2004], Auer et al. [2007], Kleinberg et al. [2008a]. Plus précisément :

(i): On propose le premier algorithme pratique pouvant s'adapter à (presque) n'importe quel espace : HOO (*Hierarchical Optimistic Optimization*).

(ii): On montre que HOO atteint la vitesse minimax (à un terme logarithmique près) dans de nombreux cas, par exemple si $\mathcal{X} = \mathbb{R}^D$ muni de la norme euclidienne et que l'on considère les fonctions moyennes 1-Lipschitz. La vitesse minimax dans ce cas est $n^{\frac{D+1}{D+2}}$ (à un terme logarithmique près).

(iii): On définit une nouvelle notion de dimension de la fonction moyenne, la *near-optimality dimension*, et on montre que le regret de HOO sur une fonction moyenne de dimension near-opt d est en fait $n^{\frac{d+1}{d+2}}$. En particulier pour le bon choix de la métrique on montre que dans beaucoup de cas intéressant $d = 0$ quelle que soit la dimension ambiante.

(iv): On montre que seul le comportement de la fonction moyenne au voisinage de son maximum est important pour obtenir les vitesses ci-dessus.

1.3. Le problème de la planification. Le bandit stochastique avec infinité de bras couvre un champ très large d'applications. En particulier il inclut le problème classique de la planification où le statisticien doit préparer un plan d'actions en ayant seulement une expérience limitée de son environnement. Dans le Chapitre 5 on développe cette correspondance et on introduit l'algorithme OLOP (Open Loop Optimistic Planning), dérivé de l'algorithme HOO du Chapitre 4. Cela nous permet d'obtenir le premier algorithme minimax optimal dans le cas d'un environnement stochastique avec récompenses actualisées.

1.4. Regret simple. Les deux derniers chapitres de la première partie concernent à nouveau le jeu du bandit stochastique avec un nombre fini de bras, mais vu sous un angle différent. Dans le jeu classique le joueur doit à la fois chercher quel bras a la meilleure récompense moyenne et dans le même temps exploiter le bras qu'il pense être le meilleur afin de minimiser son regret. Dans les chapitres 6 et 7 on s'affranchit de cette seconde contrainte et l'unique objectif du joueur

Paramètres connus : nombre de bras K ; nombre de tours n avec $n \geqslant K \geqslant 2$;

Paramètres inconnus : K distributions de probabilités $\nu_1, \ldots, \nu_K$ sur $[0,1]$ (ayant pour moyennes $\mu_1, \ldots, \mu_K$).

A chaque tour $t = 1, 2, \ldots, n$;
(1) le joueur choisit $I_t \in \{1, \ldots, K\}$;
(2) l'environnement tire la récompense $Y_t \sim \nu_{I_t}$ indépendamment du passé étant donné I_t.

A la fin du tour n le joueur choisit $J_n \in \{1, \ldots, K\}$.

But : Minimiser le regret simple $r_n = \max_{i \in \{1,\ldots,K\}} \mu_i - \mu_{J_n}$.

FIG. 4: Regret simple pour le jeu du bandit stochastique.

devient la découverte du meilleur bras. On introduit un nouveau type de regret, le regret simple, qui permet de formaliser cette idée. Le jeu correspondant est décrit dans la Figure 4. On peut noter que ce problème correspond tout simplement à l'optimisation d'une fonction stochastique sur un domaine discret.

La motivation initiale pour ce travail provient de considérations sur les problèmes réels que modélise le jeu du bandit. Par exemple dans la construction d'une intelligence artificielle pour le jeu de Go on cherche un algorithme qui étant donné une position du Goban (i.e., du plateau de jeu) va donner le meilleur coup à jouer. Pour ce faire on suppose que l'algorithme peut estimer (avec un bruit) la valeur d'un coup et qu'il dispose d'un nombre fini d'estimations. Ainsi durant le temps imparti il peut tester les différentes options pour essayer de trouver la meilleure. Clairement dans ce cas utiliser un algorithme minimisant le regret cumulé n'a pas de sens et seul le regret simple est à prendre en considération. Pourtant l'une des meilleurs I.A. actuelle est basée sur un algorithme pour le jeu du bandit classique, voir Gelly et al. [2006]. C'est dans ce contexte que se place le Chapitre 6 où l'on étudie les liens entre le regret simple et le regret cumulé. Le résultat principal est une borne inférieure sur les performances en terme de regret simple étant données celles du regret cumulé. Essentiellement le résultat peut s'énoncer sous la forme

$$\forall \nu_1, \ldots, \nu_K, \ \exists\, C > 0 : \mathbb{E} r_n \geqslant \exp(-C\, \overline{R}_n). \tag{1.1}$$

Autrement dit, meilleur est le regret cumulé, moins bon sera le regret simple. Ce résultat est vrai dans un sens *distribution-dependent*, c'est à dire quand on autorise les bornes à dépendre des distributions de probabilités sur les bras (ici sous la forme de la constante C). On montre aussi dans le Chapitre 6 qu'il est facile d'obtenir des algorithmes minimax optimaux pour le regret simple (du moins à un facteur logarithmique près), y compris pour des algorithmes ayant une vitesse de convergence (du regret simple vers 0) sous optimal en *distribution-dependent*.

Motivés par ces résultats on développe dans le Chapitre 7 deux nouveaux algorithmes spécialement conçus pour minimiser le regret simple, SR (*Successive Rejects*) et UCB-E (*Upper Confidence Bound Exploration*). On étudie leurs vitesses de convergence *distribution-dependent* et on prouve l'optimalité de ces dernières à un facteur logarithmique près. En particulier, cette analyse nous permet de prouver que pour trouver le meilleur bras, il faut un nombre de tours de jeu de l'ordre de $\sum_i 1/\Delta_i^2$, où la somme se fait sur les indices i sous-optimaux et $\Delta_i = \max_{1 \leqslant j \leqslant K} \mu_j - \mu_i$ représente la sous-optimalité du bras i. Ce résultat généralise le fait bien connu qu'il faut de l'ordre

de $1/\Delta^2$ tirages pour differencier les moyennes de deux distributions de probabilitiés ayant un écart de Δ entre leurs moyennes.

2. Les fondations du *clustering*

Le *clustering* peut être défini informellement comme la recherche de "groupes" dans un ensemble de données. En général on cherche des groupes ayant un ou plusieurs des attributs suivants.

- Les groupes permettent une représentation compréhensible par un être humain des données.
- Les groupes permettent de découvrir des catégories de données "similaires" dans notre ensemble.
- La création de ces groupes est une étape de traitement préliminaire de l'ensemble des données pour ensuite utiliser un algorithme directement sur les groupes plutôt que sur les données brut. On veut alors que ces groupes améliorent l'efficacité de l'algorithme.

Cette liste n'est pas exhaustive, mais elle est suffisante pour se rendre compte qu'un traitement théorique unifié du *clustering* est une tâche ardu.

2.1. Des algorithmes consistants. Dans le Chapitre 8 on adopte un point de vue dérivé de l'apprentissage statistique. On se donne les éléments suivants : un ensemble mesurable $\mathcal{X}$ muni d'une distribution de probabilité inconnue $\mathbb{P}$, un ensemble de données $\{X_1, \dots, X_n\}$ tirées i.i.d selon $\mathbb{P}$, un ensemble de fonctions $\mathcal{F} \subset \{1, \dots, K\}^{\mathcal{X}}$ représentant les types de groupes qu'on s'autorise, et une fonction de qualité $Q : \mathcal{F} \to \mathbb{R}^+$ qui dépend de $\mathbb{P}$. Le but est alors de construire un ensemble de groupes $f_n \in \mathcal{F}$ qui optimise notre fonction de qualité Q. Suivant la stratégie classique de l'apprentissage statistique, on construit un estimateur Q_n de Q avec notre ensemble de données et on choisit f_n qui optimise Q_n (éventuellement sur un sous-ensemble $\mathcal{F}_n \subset \mathcal{F}$). On se pose alors la question naturelle de la convergence de $Q(f_n)$ vers le véritable optimum de Q.

Notre première contribution est un théorème très général qui donne des conditions suffisantes sur $Q_n, \mathcal{F}_n, Q, \mathcal{F}$ et $\mathbb{P}$ pour que la procédure décrite précédemment soit (faiblement) consistante, i.e., $Q(f_n)$ converge (en probabilité) vers l'optimum de Q. Les résultats classiques de l'apprentissage statistique ne peuvent pas s'appliquer pour obtenir ce résultat, notamment car on doit autoriser des fonctions de qualités Q qui ne s'écrivent pas comme des espérances (i.e., $Q(f) = \mathbb{E}(\Omega(f, X))$) ainsi que des estimateurs biaisés de Q (i.e., $\mathbb{E}Q_n \neq Q$). En effet ces conditions sont la normes plutôt que l'exception dans les travaux sur le *clustering*. De plus nous autorisons $\mathcal{F}_n$ a dépendre des données. Ces hypothèses rendent les techniques classiques, tel que l'étape de symétrisation, plus difficiles à appliquer.

A partir de ce théorème nous dérivons un nouvel algorithme, NNC (*Nearest Neighbor Clustering*), qui peut être vu comme un équivalent de la méthode des plus proches voisins dans le contexte non-supervisé du *clustering*, et pour lequel nous prouvons sa consistance sur de nombreuses fonctions objectifs classiques, telles que *k-means*, *Ratio Cut*, *Normalized Cut*, etc.

2.2. La stabilité : une méthode de sélection de modèles. Dans le Chapitre 9 nous abordons une question complètement différente, comment choisir le nombre de groupes K que contient notre ensemble de données ? Pour simplifier la discussion nous nous plaçons dans un cas où on veut utiliser l'algorithme *k-means* pour construire les groupes. De plus, afin d'avoir une notion du "bon" nombre de groupes, on supposera que nos données sont générées par une mixture de Gaussiennes (sur $\mathbb{R}^d$) suffisamment ecartées. Dans ces conditions, il a été suggéré expérimentalement la procédure suivante : générer plusieurs ensembles de données et lancer l'algorithme *k-means* avec différents K. On choisit alors le nombre K pour lequel les résultats de *k-means* sont les plus

stables.

Cette procédure a été étudiée dans un cadre idéal dans une serie de papiers, [Ben-David et al., 2006, 2007, Shamir and Tishby, 2008a, Ben-David and von Luxburg, 2008, Shamir and Tishby, 2008b,c]. D'une manière générale la conclusion de ces travaux est que la stabilité ne peut pas fonctionner comme méthode de sélection du nombre de groupes (dans un cadre idéal et asymptotique).

Dans le Chapitre 9 nous nous plaçons dans le cadre "réel", et considérons l'algorithme *k-means* avec sa particularité de tomber dans des extremas locaux de la fonction objectif. Dans une première partie nous prouvons (partiellement) que la méthode de stabilité fonctionne, du moment que l'on initialise correctement l'algorithme. La seconde partie est dédiée à l'analyse d'une méthode non-triviale d'initialisation, PRUNED MINDIAM, pour laquelle nous prouvons qu'elle satisfait nos conditions pour que la méthode de stabilité fonctionne.

Part 1

Bandits Games

CHAPTER 2

Multi-Armed Bandits

This chapter is meant to be an introduction to the bandit theory. We introduce the multi-armed bandit problem in both stochastic and adversarial settings. We present different motivating examples as well as the basic theoretical results. For some of them we propose a statement with improved constants and/or simpler proofs. At the end of the chapter we discuss possible extensions of this classical problem and highlight the contributions of this thesis to some of them.

Contents

1. Bandits problems

The term *bandit* refers to the usual name of a Casino's slot machine ("one-armed bandit"). In a multi-armed bandit problem a player (or a forecaster) is facing a finite number of slot machines (or *arms*). He allocates sequentially his coins (one at time) on different machines and earns some money (its *reward*) depending on the machine he selected. His goal is simply to earn as much money as possible. As we will see below, the most important feature of this model is the assumption on the slot machines' reward generation process.

1.1. Stochastic multi-armed bandit. In its original formulation, Robbins [1952], each arm corresponds to an unknown probability distribution on $[0, 1]$. At each time step $t \in \mathbb{N}$ the forecaster selects (or *pulls*) one arm I_t, and then he receives a reward Y_t sampled from the distribution corresponding to the selected arm and independently from the past given that arm. The forecaster's goal is to maximize the sum of rewards $\sum_{t=1}^{n} Y_t$ where $n \in \mathbb{N}$ is the time horizon. The forecaster does not necessarily know in advance the time horizon and in that case we say that his strategy is *anytime*.

Known parameters: number of arms K; number of rounds n with $n \geqslant K \geqslant 2$.

Unknown parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$.

For each round $t = 1, 2, \dots, n$

(1) the forecaster chooses $I_t \in \{1, \dots, K\}$;

(2) the environment draws the reward $Y_t \sim \nu_{I_t}$ independently from the past given I_t.

Goal: Maximize the cumulative gain $\sum_{t=1}^{n} Y_t$.

Figure 1: Stochastic multi-armed bandit game.

If the distributions were known, one would always pull the arm with the highest mean reward in order to maximize the cumulative rewards. To analyze the behavior of a forecaster we compare its performance with this optimal strategy. In other terms we study the *regret* of the forecaster for not playing optimally. Let $K \geqslant 2$ be the number of arms and for $i \in \{1, \dots, K\}$ we note ν_i the probability distribution of arm i and μ_i its mean. We also set $\mu^* = \max_{i \in \{1,\dots,K\}} \mu_i$ and $i^* \in \operatorname{argmax}_{i \in \{1,\dots,K\}} \mu_i$. Then the cumulative pseudo-regret (this terminology will be explained in Section 1.2) of the forecaster is:

$$\overline{R}_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} \mu_{I_t} \tag{2.1}$$

where the expectation is taken with respect to the random drawing of the rewards (which influence the sequence $I_1, \dots, I_n$). It represents the average regret of the forecaster with respect to the best arm on average. In the following we refer to this problem as the stochastic multi-armed bandit, see Figure 1 for a summary.

The historical motivation for this model was given by medical trials, Thompson [1933]. The setting is as follows. A set of K drugs is at disposal to cure one disease, and patients are sequentially presented without any side information. We assume that the effectiveness of each drug is the same for any patient. More formally the actual success of a drug on a patient is a Bernoulli random variable whose parameter depends solely on the drug. The objective is to maximize the number of healed patients, or more precisely to perform almost as well as the best drug. In particular we do not want to use suboptimal treatments too often but rather focus as soon as possible on the best treatment. Thus a trade-off appears between the exploration of the different drugs to estimate their performance and the exploitation of the results we have obtained so far to be the more efficient. This is in fact one of the basic motivation for the stochastic multi-armed bandit game, it is the simplest model in which an exploration/exploitation dilemma appears.

1.2. Adversarial multi-armed bandit. To motivate the adversarial bandit let us consider again the initial example of gambling on slot machines. We assume now that we are in rigged casino, where the owner (called the *adversary*) sets for each $i \in \{1, \dots, K\}$ the sequence of gains $(g_{i,t})_{1 \leqslant t \leqslant n} \in [0,1]^n$ for slot machine i. Note that it is not in the interest of the owner to simply set all the gains to zero, since in that case one will eventually go to another casino! Now recall that a forecaster selects sequentially one arm $I_t \in \{1, \dots, K\}$ at each time step $1 \leqslant t \leqslant n$ and observes (and earns) the gain $g_{I_t,t}$. Is it still possible to be competitive in such setting ? As a first step, and by analogy with the stochastic case, one considers the best single arm in hindsight and

seek for a forecaster which earns almost as much rewards as this arm. More precisely we define the cumulative regret of the forecaster as:

$$R_n = \max_{i\in\{1,\dots,K\}} \sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t}. \tag{2.2}$$

In the particular example of the rigged casino we say that the adversary is *oblivious*, because the opponent's strategy (represented by the sequence of gains) is oblivious to our actions. In a more general framework, the adversary may depend on our actions, in that case the adversary is *non-oblivious*. For instance the Casino's owner may look at the forecaster's strategy to design even more evil sequences of gains. See Figure 2 for a summary of the general problem. Note that for a non-oblivious adversary the interpretation of the regret is tricky and may be seen as unnatural. In particular we can not evaluate the cumulative gain that we would have obtained by playing a single arm for the n rounds since the only information available on the adversary is the sequence of gain vectors that was produced against the forecaster's strategy. Thus the regret can only be computed on this sequence of gains.

In this setting the goal is to obtain bounds in high probability or in expectation (with respect to both eventual randomization of the forecaster and the adversary) on the regret for *any* opponent. In the case of a non-oblivious opponent this is not an easy task, and we usually bound first the pseudo-regret:

$$\overline{R}_n = \max_{i\in\{1,\dots,K\}} \mathbb{E}\sum_{t=1}^{n} g_{i,t} - \mathbb{E}\sum_{t=1}^{n} g_{I_t,t}. \tag{2.3}$$

which compares the forecaster's average regret with respect to the best arm arm on average. Remark that the randomization of the adversary is not very important since we ask for bounds which hold for any opponent. However, by allowing this randomization, we recover the stochastic bandit as a special case of the adversarial bandit, in particular for stochastic bandits equations (2.1) and (2.3) coincide. On the other hand it is fundamental to allow randomization for the forecaster. Indeed, given a deterministic forecaster, the oblivious adversary defined as follows

$$\begin{cases} \text{if } I_t \neq 1,\ g_{1,t} = 1 \text{ and } g_{i,t} = 0 \text{ for all } i \neq 1; \\ \text{if } I_t = 1,\ g_{2,t} = 1 \text{ and } g_{i,t} = 0 \text{ for all } i \neq 2; \end{cases}$$

impose $R_n \geqslant n/2$. Note that this adversary is allowed (and oblivious) because the forecaster's strategy is deterministic.

1.3. Modern motivating examples. While the historical motivation for bandits was medical trials, it has been realized in the past decades that bandits games model a number of more sophisticated and relevant applications. We describe here six examples, ranging from theoretical to applied, where bandits algorithms have been used or are currently under investigation. Note that most of these examples are better modeled by extensions of the basic game, see Section 4.

(1) **Online learning with expert advice and limited feedback:** The traditional statistical point of view on learning is the following. The learner (or forecaster) has access to a data set $(\xi_t)_{1\leqslant t\leqslant n} \in \Xi^n$ and has to output a prediction $f_n \in \mathcal{F}$ where $\mathcal{F}$ is a class of functions usually defined by computational and statistical modeling considerations. This prediction is evaluated by a loss function $L : \mathcal{F} \to \mathbb{R}^+$, and the forecaster's goal is to minimize the expected loss. A common example fitting in this framework is the

Known parameters: number of arms K; number of rounds n with $n \geqslant K \geqslant 2$.

For each round $t = 1, 2, \ldots, n$

(1) the forecaster chooses $I_t \in \{1, \ldots, K\}$, eventually with the help of an external randomization;

(2) simultaneously the adversary selects a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0, 1]^K$, eventually with the help of an external randomization;

(3) the forecaster receives (and observes) the reward $g_{I_t,t}$. He does not observe the gains of the other arms.

Goal: Maximize the cumulative gain $\sum_{t=1}^{n} g_{I_t,t}$.

Figure 2: Adversarial multi-armed bandit game.

one of pattern classification, where the forecaster faces pairs $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, drawn i.i.d. from an unknown probability distribution $\mathbb{P}$. A prediction is a classifier, that is, $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ and the loss of a classifier is the expected number of mistakes, that is, $L(f) = \mathbb{P}(f(X) \neq Y)$. A limitation of this viewpoint is that the data set is fixed once and for all. This assumption is not valid for many of the modern problems in statistics. Indeed, whether we consider the Internet network, consumer data sets, or financial market, a common feature emerges: the data are dynamic and continuously evolving. The online learning model addresses this issue. Here the forecaster faces the data sequentially, and at each time step a prediction has to be made. This prediction incurs an instant loss given by a function $\ell : \mathcal{F} \times \Xi \to \mathbb{R}^+$. More precisely at time step $t \in \mathbb{N}$ the learner already observed $\xi_1, \ldots, \xi_{t-1}$ and made the predictions $f_1, \ldots, f_{t-1}$. Then the learner makes the prediction f_t, receives a new data point ξ_t and suffers the loss $\ell(f_t, \xi_t)$. Thus the cumulative loss of the forecaster after n steps is $L_n = \sum_{t=1}^{n} \ell(f_t, \xi_t)$.

One sub-model of this framework has attracted a lot of attention in the last years, namely the one of online learning with expert advice: A set of K experts are playing the online learning game described above and the forecaster selects at each time step which expert's prediction to follow, based on its past experience with the different experts. This problem is now very well understood, see Cesa-Bianchi and Lugosi [2006]. One interesting and challenging extension is the model of online learning with limited feedback. Here the learner only observes the suffered loss $\ell(f_t, \xi_t)$ rather than the data point ξ_t. This model can be viewed as a bandit game, where the set of experts correspond to the set of arms.

(2) **Black-box stochastic optimization:** Consider an algorithm with a parameter to tune, and whose performance can easily be evaluated. Assume that either the algorithm has some internal stochasticity, or that it can only be evaluated with an additional noise. By viewing this problem as a stochastic bandit game, where the set of arms corresponds to the set of possible parameter values, we obtain strategies to automatically tune the parameter over time. We refer the reader to Chapter 4 for more details on this approach.

(3) **Ads placement on a web-page:** With the pay-per-click model, each click on an ad proposed by a search engine earns some revenue. As users arrive, the search engine can display ads sequentially in order to maximize the total revenue. This problem can be cast as a bandit game, with the set of arms being the set of ads and each user corresponds to a time step. Recent works to apply bandits in this framework include Pandey et al.

[2007b], Pandey et al. [2007a], Liu and Zhao [2008] and Chakrabarti et al. [2009], see also Section 4.3 for the bandit with side information and the references therein.

(4) **Packets routing:** Consider a network represented by a graph in which one has to send a sequence of packets from one vertex to another. For each packet one chooses a path through the graph and suffers a certain delay. Depending on the traffic, the delays on the edges may change and the only information available is the delay on the path chosen at a given time step. The goal is to minimize the total delay for the sequence of packets. This problem can be cast as a bandit game, with the set of arms being the set of paths between the two vertices and each packet corresponds to a time step. Many works have been done for this particular problem, including Awerbuch and Kleinberg [2004], McMahan and Blum [2004], György et al. [2007], see also the references in Section 4.1 for combinatorial bandits and linear bandit optimization.

(5) **Tree-search:** A recent successful application of bandits algorithms is the MoGo program of Gelly et al. [2006] which plays computer Go at a world-class level. With a Monte-Carlo method to evaluate the value of a position it is possible to see the Go problem as a hierarchy of bandits. The authors address this problem with the UCT strategy of Kocsis and Szepesvari [2006] which is derived from the UCB strategy, see Section 2.2.

(6) **Channel allocation for cellphones:** During a communication between two cellphones, the operator may change the channel several times. Here the set of arms corresponds to the set of possible channels and a time step represents a time interval where the channel is fixed. Opportunistic communication systems rely on the same idea and has been recently studied in the more general framework of partially observable Markov decision process in Zhao et al. [2005], Liu and Zhao [2008], Filippi et al. [2009].

2. Upper bounds on the cumulative regret

In this section we propose different forecasters both for the stochastic and adversarial bandit game. We analyze the performance of the algorithms by proving upper bounds on their regrets. In fact, we prove three different type of bounds, (i) on the pseudo-regret, (ii) on the expected regret, and (iii) bounds on the regret which hold with high probability. The weakest statement and also the easier to obtain is (i). In sections 2.1 and 2.2 we focus on this quantity. Then in Section 2.3 we show how to extend these results to (ii) and (iii).

2.1. Pseudo-regret in adversarial bandits. As we said in Section 1.2, it is necessary to consider randomized forecaster to obtain non-trivial guarantee on the cumulative regret in the adversarial bandit game. We describe here the randomized forecaster Exp3 of Auer et al. [2003] which is based on two fundamental ideas. The first one is that, despite the fact that only the gain for one arm is observed, it is still possible to build an unbiased estimator of the gain for any arm with a simple trick. Namely, if the next arm I_t to be played is drawn from a probability distribution $p_t = (p_{1,t}, \ldots, p_{K,t})$, then $\widetilde{g}_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$ is an unbiased estimator (with respect to the drawing of I_t) of $g_{i,t}$. The second idea, which goes back to Littlestone and Warmuth [1994], is to use an exponential reweighting of the cumulative estimated gain to define the new probability distribution over the arm.

In recent works, Cesa-Bianchi and Lugosi [2006], this strategy has been described in a loss setting, that is instead of maximizing the cumulative gain the goal is to minimize the cumulative loss. With a linear transformation on the rewards we can reduce a gain game to a loss game and

Exp3 (Exponential weights algorithm for Exploration and Exploitation) without mixing:

Parameter: a non-increasing sequence of real numbers $(\eta_t)_{t\in\mathbb{N}}$.

Let p_1 be the uniform distribution over $\{1,\dots,K\}$.

For each round $t = 1, 2, \dots, n$

(1) Draw an arm I_t from the probability distribution p_t.

(2) Compute the estimated loss for each arm: $\widetilde{\ell}_{i,t} = \frac{1-g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$ and update the estimated cumulative loss: $\widetilde{L}_{i,t} = \sum_{s=1}^{t}\widetilde{\ell}_{i,s}$.

(3) Compute the new probability distribution over the arms $p_{t+1} = (p_{1,t+1},\dots,p_{K,t+1})$ where:

$$p_{i,t+1} = \frac{\exp\left(-\eta_t\widetilde{L}_{i,t}\right)}{\sum_{k=1}^{K}\exp\left(-\eta_t\widetilde{L}_{k,t}\right)}.$$

Figure 3: Exp3 forecaster.

use the Exp3 strategy for losses. This way we obtain better bounds in terms of the constants [1]. Moreover this modification has an other advantage, in the initial formulation the authors had to consider a mixture of exponential weights with the uniform distribution on the set of arms, but as it was already noted in Stoltz [2005], this mixing is not necessary to have bounds on the pseudo-regret when Exp3 is working on losses [2]. Thus we can focus here on Exp3 without mixing, see Figure 3.

We provide two different bounds. In (2.5) we assume that the forecaster does not know the number of rounds n, that is we consider the anytime version of the algorithm. On the other hand, in (2.4) we prove that we can achieve a better constant with the knowledge of the time horizon. Moreover, we improve these constants with respect to previous results. In Auer et al. [2003] equation (2.4) was proved with a constant $2\sqrt{e-1}$ for the algorithm with mixing and gains, while in Stoltz [2005] the constant was $2\sqrt{2}$ without mixing and with losses. Here we provide a constant $\sqrt{2}$ for the algorithm without mixing and working with linear transformations of the gains.

THEOREM 2.1 (Pseudo-regret of Exp3). *Exp3 without mixing and with $\eta_t = \eta = \sqrt{\frac{2\log K}{nK}}$ satisfies:*

$$\overline{R}_n \leqslant \sqrt{2nK\log K}. \tag{2.4}$$

On the other hand with $\eta_t = \sqrt{\frac{2\log K}{tK}}$ it satisfies:

$$\overline{R}_n \leqslant \frac{3}{\sqrt{2}}\sqrt{nK\log K}. \tag{2.5}$$

PROOF. We will prove that for any non-increasing sequence $(\eta_t)_{t\in\mathbb{N}}$, Exp3 without mixing satisfies:

$$\overline{R}_n \leqslant \frac{K}{2}\sum_{t=1}^{n}\eta_t + \frac{\log K}{\eta_n}. \tag{2.6}$$

[1] The technical reason is that when Exp3 works on losses, it uses the exponential function on negative real numbers where it enjoys a simple upper bound by a quadratic function, namely $\exp(x) \leqslant 1 + x + x^2/2$.

[2] The mixing is necessary when working with gains to ensure that Exp3 is working with the exponential function on the interval $[0,1]$, where it can also be easily bounded by a quadratic function, namely $\exp(x) \leqslant 1 + x + x^2$.

Equation (2.4) then trivially follows from (2.6). On the other hand for (2.5) we use (2.6) and

$$\sum_{t=1}^{n} \frac{1}{\sqrt{t}} \leqslant \int_0^n \frac{1}{\sqrt{t}} dt = 2\sqrt{n}.$$

We provide now a proof of (2.6) in five steps.

First step: Useful equalities.
The following equalities can be verified very easily:
(2.7)

$$\mathbb{E}_{i\sim p_t} \widetilde{\ell}_{i,t} = 1 - g_{I_t,t}; \qquad \mathbb{E}_{I_t\sim p_t} \widetilde{\ell}_{i,t} = 1 - g_{i,t}; \qquad \mathbb{E}_{i\sim p_t} \widetilde{\ell}^2_{i,t} = \frac{(1-g_{I_t,t})^2}{p_{I_t,t}}; \qquad \mathbb{E}_{I_t\sim p_t} \frac{1}{p_{I_t,t}} = K.$$

In particular this implies

$$\sum_{t=1}^{n} g_{k,t} - \sum_{t=1}^{n} g_{I_t,t} = \sum_{t=1}^{n} \mathbb{E}_{i\sim p_t} \widetilde{\ell}_{i,t} - \sum_{t=1}^{n} \mathbb{E}_{I_t\sim p_t} \widetilde{\ell}_{k,t}. \tag{2.8}$$

The key step in the proof is now to consider log-moment of $\widetilde{\ell}_{i,t}$:

$$\mathbb{E}_{i\sim p_t} \widetilde{\ell}_{i,t} = \frac{1}{\eta_t} \log \mathbb{E}_{i\sim p_t} \exp\left(-\eta_t(\widetilde{\ell}_{i,t} - \mathbb{E}_{k\sim p_t} \widetilde{\ell}_{k,t})\right) - \frac{1}{\eta_t} \log \mathbb{E}_{i\sim p_t} \exp\left(-\eta_t \widetilde{\ell}_{i,t}\right). \tag{2.9}$$

In the next two steps we study the two terms of the right hand side in (2.9).

Second step: Study of the first term in (2.9)**.**
We use the inequalities $\log x \leqslant x - 1$ and $\exp(-x) - 1 + x \leqslant x^2/2, \forall x \geqslant 0$:

$$\begin{aligned}
\log \mathbb{E}_{i\sim p_t} \exp\left(-\eta_t(\widetilde{\ell}_{i,t} - \mathbb{E}_{k\sim p_t} \widetilde{\ell}_{k,t})\right) &= \log \mathbb{E}_{i\sim p_t} \exp\left(-\eta_t \widetilde{\ell}_{i,t}\right) + \eta_t \mathbb{E}_{k\sim p_t} \widetilde{\ell}_{k,t} \\
&\leqslant \mathbb{E}_{i\sim p_t} \left(\exp\left(-\eta_t \widetilde{\ell}_{i,t}\right) - 1 + \eta_t \widetilde{\ell}_{i,t}\right) \\
&\leqslant \mathbb{E}_{i\sim p_t} \eta_t^2 \widetilde{\ell}^2_{i,t}/2 \\
&\leqslant \frac{\eta_t^2}{2p_{I_t,t}} \qquad (2.10)
\end{aligned}$$

where the last step comes from the third equality in (2.7).

Third step: Study of the second term in (2.9)**.**
Let $\widetilde{L}_{i,0} = 0$, $\Phi_0(\eta) = 0$ and $\Phi_t(\eta) = \frac{1}{\eta} \log \frac{1}{K} \sum_{i=1}^K \exp\left(-\eta \widetilde{L}_{i,t}\right)$. Then by definition of p_t we have:

$$\begin{aligned}
-\frac{1}{\eta_t} \log \mathbb{E}_{i\sim p_t} \exp\left(-\eta_t \widetilde{\ell}_{i,t}\right) &= -\frac{1}{\eta_t} \log \frac{\sum_{i=1}^K \exp\left(-\eta_t \widetilde{L}_{i,t}\right)}{\sum_{i=1}^K \exp\left(-\eta_t \widetilde{L}_{i,t-1}\right)} \\
&= \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t). \qquad (2.11)
\end{aligned}$$

Fourth step: Summing.
Putting together (2.8), (2.9), (2.10) and (2.11) we obtain

$$\sum_{t=1}^{n} g_{k,t} - \sum_{t=1}^{n} g_{I_t,t} \leqslant \sum_{t=1}^{n} \frac{\eta_t}{2p_{I_t,t}} + \sum_{t=1}^{n} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) - \sum_{t=1}^{n} \mathbb{E}_{I_t\sim p_t} \widetilde{\ell}_{k,t}.$$

The first term is easy to bound in expectation since by the tower rule and the last equality in (2.7) we have:

$$\mathbb{E}\sum_{t=1}^{n}\frac{\eta_t}{2p_{I_t,t}} = \mathbb{E}\sum_{t=1}^{n}\mathbb{E}_{I_t\sim p_t}\frac{\eta_t}{2p_{I_t,t}} = \frac{K}{2}\sum_{t=1}^{n}\eta_t.$$

For the second term we start with an Abel transformation:

$$\sum_{t=1}^{n}\left(\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t)\right) = \sum_{t=1}^{n-1}\left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right) - \Phi_n(\eta_n)$$

since $\Phi_0(\eta_1) = 0$. Remark that

$$\begin{aligned}-\Phi_n(\eta_n) = \frac{\log K}{\eta_n} - \frac{1}{\eta_n}\log\left(\sum_{i=1}^{K}\exp\left(-\eta_n\widetilde{L}_{i,n}\right)\right) &\leqslant \frac{\log K}{\eta_n} - \frac{1}{\eta_n}\log\left(\exp\left(-\eta_n\widetilde{L}_{k,n}\right)\right)\\ &= \frac{\log K}{\eta_n} + \sum_{t=1}^{n}\widetilde{\ell}_{k,t}.\end{aligned}$$

Thus

$$\mathbb{E}\left(\sum_{t=1}^{n}g_{k,t} - \sum_{t=1}^{n}g_{I_t,t}\right) \leqslant \frac{K}{2}\sum_{t=1}^{n}\eta_t + \frac{\log K}{\eta_n} + \mathbb{E}\sum_{t=1}^{n-1}\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t).$$

To conclude the proof we show that $\Phi_t'(\eta) \geqslant 0$ and thus since $\eta_{t+1} \leqslant \eta_t$ we obtain $\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t) \leqslant 0$. Let π be the uniform distribution over $\{1,\dots,K\}$ and $p_{i,t}^{\eta} = \frac{\exp\left(-\eta\widetilde{L}_{i,t}\right)}{\sum_{k=1}^{K}\exp\left(-\eta\widetilde{L}_{k,t}\right)}$, then:

$$\begin{aligned}\Phi_t'(\eta) &= -\frac{1}{\eta^2}\log\left(\frac{1}{K}\sum_{i=1}^{K}\exp\left(-\eta\widetilde{L}_{i,t}\right)\right) - \frac{1}{\eta}\frac{\sum_{i=1}^{K}\widetilde{L}_{i,t}\exp\left(-\eta\widetilde{L}_{i,t}\right)}{\sum_{i=1}^{K}\exp\left(-\eta\widetilde{L}_{i,t}\right)}\\ &= \frac{1}{\eta^2}\frac{1}{\sum_{i=1}^{K}\exp\left(-\eta\widetilde{L}_{i,t}\right)}\sum_{i=1}^{K}\exp\left(-\eta\widetilde{L}_{i,t}\right)\times\left(-\eta\widetilde{L}_{i,t} - \log\left(\frac{1}{K}\sum_{i=1}^{K}\exp\left(-\eta\widetilde{L}_{i,t}\right)\right)\right)\\ &= \frac{1}{\eta^2}\sum_{i=1}^{K}p_{i,t}^{\eta}\log(Kp_{i,t}^{\eta})\\ &= \frac{1}{\eta^2}\mathrm{KL}(p_t^{\eta},\pi) \geqslant 0.\end{aligned}$$

□

2.2. Pseudo-regret in stochastic bandits. In this section, we describe a simple modification proposed by Audibert et al. [2009] of the UCB1 policy of Auer et al. [2002], and the UCB-V policy of Audibert et al. [2009]. The main idea for both strategies is to measure the performance of each arm by an index, and at each round, the forecaster chooses the arm having the highest index (see Figure 4). This index is meant to be an upper confidence bound on the mean reward which holds with high probability. This idea can be traced back to Agrawal [1995a]. For instance the fact that UCB's index is a high probability bound on the true mean is a consequence of Theorem 10.1. For UCB-V it follows from Theorem 10.3. We only present the results for the anytime formulation of both strategies, see Audibert et al. [2009] for UCB with known time horizon.

To describe properly the strategies in Figure 4 we need to introduce a few notations. It is common for stochastic bandits games to introduce the random variable $X_{i,s}$ which represents the gain obtained while pulling arm i for the s^{th} time. In particular the law of $X_{i,s}$ is ν_i. We recall that

UCB (Upper Confidence Bound), UCB-V (Upper Confidence Bound with Variance):

Parameter: exploration rate $\alpha > 0$.

For an arm i, define its index $B_{i,s,t}$ by

$$\text{UCB index:} \qquad B_{i,s,t} = \widehat{\mu}_{i,s} + \sqrt{\frac{\alpha \log(t)}{s}},$$

$$\text{UCB-V index:} \qquad B_{i,s,t} = \widehat{\mu}_{i,s} + \sqrt{\frac{2\alpha V_{i,s} \log(t)}{s}} + 3\alpha \frac{\log(t)}{s}.$$

for $s, t \geqslant 1$ and $B_{i,0,t} = +\infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1),t}$.

Figure 4: UCB and UCB-V policies.

μ_i is the mean of ν_i and we denote by σ_i^2 its variance. Now let $\widehat{\mu}_{i,s} = \frac{1}{s}\sum_{t=1}^{s} X_{i,t}$ (respectively $V_{i,s} = \frac{1}{s}\sum_{t=1}^{s}(X_{i,t} - \widehat{\mu}_{i,s})^2$) be the empirical mean (respectively the empirical variance) of arm i after s pulls of this arm. Let $T_i(s) = \sum_{t=1}^{s} \mathbb{1}_{I_t=i}$ denote the number of times we have drawn arm i on the first s rounds. To state the results we define the suboptimality of an arm i as $\Delta_i = \mu^* - \mu_i$. Thus the pseudo-regret can be written as:

$$\overline{R}_n = \sum_{i=1}^{K} \Delta_i \mathbb{E} T_i(n). \tag{2.12}$$

We propose two different bounds on the pseudo-regret of UCB. Equation (2.13) is a distribution-dependent bound while (2.14) is a distribution-free bound. In the former we have a logarithmic dependency on the number of rounds while in the latter it worsens to the square root of the number of rounds. As we will see in Section 3, this is not artificial and both bounds are almost optimal. Note that we improve the analysis of UCB with respect to previous works. Indeed we prove that we can take $\alpha > 1/2$, while in Audibert et al. [2009] they had to use $\alpha > 1$ (and the proof technique of Auer et al. [2002] also only works for $\alpha > 1$). The technical idea to obtain this improved result is actually borrowed from Audibert et al. [2009] [3].

THEOREM 2.2 (Pseudo-regret of UCB). *In the stochastic bandit game, UCB with $\alpha > 1/2$ satisfies:* [4]

$$\overline{R}_n \leqslant \sum_{i:\Delta_i>0} \frac{4\alpha}{\Delta_i} \log(n) + \Delta_i \left(1 + \frac{4}{\log(\alpha + 1/2)} \left(\frac{\alpha + 1/2}{\alpha - 1/2}\right)^2\right), \tag{2.13}$$

and

$$\overline{R}_n \leqslant \sqrt{nK \left(4\alpha \log n + 1 + \frac{4}{\log(\alpha + 1/2)} \left(\frac{\alpha + 1/2}{\alpha - 1/2}\right)^2\right)}. \tag{2.14}$$

[3] In fact one can propose another improvement of the analysis and show that the leading constant is of order of $(1/\sqrt{2} + \sqrt{\alpha})^2$ rather than 4α. To do this one has to introduce an additional free variable $c \leqslant 1$ as a multiplicative factor of the confidence term in (2.17). However this makes the final bound less readable and does not improve the constant in the interesting regime where α tends to $1/2$.

[4] Note that the additive constant is not optimized, see Audibert et al. [2009] for ideas to improve the analysis for that matter.

PROOF. Both proofs of (2.13) and (2.14) rely on bounding the expected number of pulls for a suboptimal arm. More precisely, in the first three steps of the proof we shall focus on proving that, for any i such that $\Delta_i > 0$,

$$\mathbb{E}T_i(n) \leqslant \frac{4\alpha \log(n)}{\Delta_i^2} + 1 + \frac{4}{\log(\alpha + 1/2)} \left(\frac{\alpha + 1/2}{\alpha - 1/2} \right)^2. \tag{2.15}$$

For ease of notations we introduce $\beta = \frac{1}{\alpha+1/2}$ and $u = \lceil \frac{4\alpha \log n}{\Delta_i^2} \rceil$. Note that with these notations, up to rounding, (2.15) is equivalent to $\mathbb{E}T_i(n) \leqslant u + \frac{4}{\log(1/\beta)(2\beta\alpha - 1)^2}$.

First step.
We show that if $I_t = i$, then it means that one the three following equations is true:

$$B_{i^*, T_{i^*}(t-1), t} \leqslant \mu^*, \tag{2.16}$$

or

$$\widehat{\mu}_{i, T_i(t-1)} > \mu_i + \sqrt{\frac{\alpha \log t}{T_i(t-1)}}, \tag{2.17}$$

or

$$T_i(t-1) < \frac{4\alpha \log n}{\Delta_i^2}. \tag{2.18}$$

Indeed, let us assume that the three equations are false, then we have:

$$B_{i^*, T_i(t-1), t} > \mu^* = \mu_i + \Delta_i \geqslant \mu_i + 2\sqrt{\frac{\alpha \log t}{T_i(t-1)}} \geqslant \widehat{\mu}_{i, T_i(t-1)} + \sqrt{\frac{\alpha \log t}{T_i(t-1)}} = B_{i, T_i(t-1), t},$$

which implies in particular that $I_t \neq i$.

Second step.
Here we bound the probability that (2.16) or (2.17) hold true. We use a peeling argument together with Hoeffding's maximal inequality. We recall that the latter is an easy consequence of Hoeffding-Azuma inequality for martingales (see Theorem 10.1) which states that for centered i.i.d random variables $X_1, X_2, \ldots$, and for any $x > 0$, $t \geqslant 1$,

$$\mathbb{P}\left(\exists s \in \{1, \ldots, t\}, \sum_{t=1}^{s} X_t > x \right) \leqslant \exp\left(-\frac{2x^2}{t} \right).$$

Now note that

$$\begin{aligned} \mathbb{P}(\text{(2.16) is true}) &\leqslant \mathbb{P}\left(\exists s \in \{1, \ldots, t\} : \widehat{\mu}_{i^*, s} + \sqrt{\frac{\alpha \log t}{s}} \leqslant \mu^* \right) \\ &= \mathbb{P}\left(\exists s \in \{1, \ldots, t\} : \sum_{\ell=1}^{s} (X_{i^*, \ell} - \mu^*) \leqslant -\sqrt{\alpha s \log(t)} \right). \end{aligned}$$

We apply the peeling argument with a geometric grid over the time interval $[1, t]$. More precisely, since $\beta \in (0, 1)$, we note that if $s \in \{1, \ldots, t\}$ then $\exists\, j \in \left\{ 0, \ldots, \frac{\log t}{\log 1/\beta} \right\} : \beta^{j+1} t < s \leqslant \beta^j t$. Thus we get

$$\mathbb{P}(\text{(2.16) is true}) \leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1} t < s \leqslant \beta^j t, \sum_{\ell=1}^{s} (X_{i^*, \ell} - \mu^*) \leqslant -\sqrt{\alpha s \log(t)} \right)$$

$$\leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \mathbb{P}\left(\exists s : \beta^{j+1}t < s \leqslant \beta^j t, \sum_{\ell=1}^{s}(X_{i^*,\ell} - \mu^*) \leqslant -\sqrt{\alpha\beta^{j+1}t\log(t)}\right).$$

To bound this last term we use Hoeffding's maximal inequality, which finally gives:

$$\mathbb{P}((2.16) \text{ is true}) \leqslant \sum_{j=0}^{\frac{\log t}{\log 1/\beta}} \exp\left(-\frac{2(\sqrt{\beta^{j+1}t\alpha\log t})^2}{\beta^j t}\right) \leqslant \left(\frac{\log t}{\log 1/\beta} + 1\right)\frac{1}{t^{2\beta\alpha}}.$$

Using the same arguments, one can prove

$$\mathbb{P}((2.17) \text{ is true}) \leqslant \left(\frac{\log t}{\log 1/\beta} + 1\right)\frac{1}{t^{2\beta\alpha}}.$$

Third step.
Using the first step we obtain:

$$\begin{aligned}
\mathbb{E}T_i(n) = \mathbb{E}\sum_{t=1}^{n}\mathbb{1}_{I_t=i} &\leqslant u + \mathbb{E}\sum_{t=u+1}^{n}\mathbb{1}_{I_t=i \text{ and } (2.18) \text{ is false}} \\
&\leqslant u + \mathbb{E}\sum_{t=u+1}^{n}\mathbb{1}_{(2.16) \text{ or } (2.17) \text{ is true}} \\
&= u + \sum_{t=u+1}^{n}\mathbb{P}((2.16) \text{ is true}) + \mathbb{P}((2.17) \text{ is true}).
\end{aligned}$$

Now using the second step we get:

$$\begin{aligned}
\sum_{t=u+1}^{n}\mathbb{P}((2.16) \text{ is true}) + \mathbb{P}((2.17) \text{ is true}) &\leqslant 2\sum_{t=u+1}^{n}\left(\frac{\log t}{\log 1/\beta} + 1\right)\frac{1}{t^{2\beta\alpha}} \\
&\leqslant 2\int_1^{+\infty}\left(\frac{\log t}{\log 1/\beta} + 1\right)\frac{1}{t^{2\beta c^2\alpha}}dt \\
&\leqslant \frac{4}{\log(1/\beta)(2\beta\alpha - 1)^2},
\end{aligned}$$

where we used a simple integration by parts for the last inequality. This concludes the proof of (2.15).

Fourth step.
Using (2.12) and (2.15) we directly obtain (2.13). On the other hand for (2.14) we write:

$$\begin{aligned}
\overline{R}_n &= \sum_{i:\Delta_i>0}\Delta_i\sqrt{\mathbb{E}T_i(n)}\sqrt{\mathbb{E}T_i(n)} \\
&\leqslant \sum_{i:\Delta_i>0}\sqrt{\mathbb{E}T_i(n)}\sqrt{4\alpha\log(n) + 1 + \frac{4}{\log(\alpha+1/2)}\left(\frac{\alpha+1/2}{\alpha-1/2}\right)^2} \\
&\leqslant K\sqrt{\frac{1}{K}\sum_{i:\Delta_i>0}\mathbb{E}T_i(n)}\sqrt{4\alpha\log(n) + 1 + \frac{4}{\log(\alpha+1/2)}\left(\frac{\alpha+1/2}{\alpha-1/2}\right)^2}
\end{aligned}$$

$$\leqslant \sqrt{nK\left(4\alpha\log(n)+1+\frac{4}{\log(\alpha+1/2)}\left(\frac{\alpha+1/2}{\alpha-1/2}\right)^2\right)},$$

where we used the concavity of the square root for the second step and $\sum_{i=1}^K T_i(n) = n$ for the last step. □

THEOREM 2.3 (Pseudo-regret of UCB-V). *In the stochastic bandit game, UCB-V with $\alpha > 1$ satisfies:* [5]

$$\overline{R}_n \leqslant 8\alpha \sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+2\right)\log(n)+\Delta_i\left(2+\frac{12}{\log(\alpha+1)}\left(\frac{\alpha+1}{\alpha-1}\right)^2\right). \tag{2.19}$$

PROOF. The proof follows the same scheme than the one of Theorem 2.2.

First step.
We show that if $I_t = i$, then it means that one the four following equations is true:

$$B_{i^*,T_{i^*}(t-1),t} \leqslant \mu^*, \tag{2.20}$$

or

$$\widehat{\mu}_{i,T_i(t-1)} > \mu_i + \sqrt{\frac{2\alpha V_{i,T_i(t-1)}\log t}{T_i(t-1)}} + 3\alpha\frac{\log t}{T_i(t-1)}, \tag{2.21}$$

or

$$V_{i,T_i(t-1)} > \sigma_i^2 + \Delta_i/2, \tag{2.22}$$

or

$$T_i(t-1) < \frac{8\alpha(\sigma_i^2+2\Delta_i)\log n}{\Delta_i^2}. \tag{2.23}$$

Indeed, let us assume that the four equations are false. We start with the following computations:

$$\sqrt{\frac{2\alpha V_{i,T_i(t-1)}\log t}{T_i(t-1)}} + 3\alpha\frac{\log t}{T_i(t-1)} \leqslant \sqrt{\frac{2\sigma_i^2+\Delta_i}{2\sigma_i^2+4\Delta_i}}\frac{\Delta_i}{2} + \frac{3\Delta_i}{4\sigma_i^2+8\Delta_i}\frac{\Delta_i}{2} \leqslant \frac{\Delta_i}{2},$$

where we used $x + \frac{1-x^2}{2} \leqslant 1$ for the last inequality. Then we have:

$$\begin{aligned} B_{i^*,T_i(t-1),t} > \mu^* = \mu_i + \Delta_i &\geqslant \mu_i + 2\left(\sqrt{\frac{2\alpha V_{i,T_i(t-1)}\log t}{T_i(t-1)}} + 3\alpha\frac{\log t}{T_i(t-1)}\right) \\ &\geqslant \widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{2\alpha V_{i,T_i(t-1)}\log t}{T_i(t-1)}} + 3\alpha\frac{\log t}{T_i(t-1)} \\ &= B_{i,T_i(t-1),t}, \end{aligned}$$

which implies in particular that $I_t \neq i$.

Second step.
Using the same reasoning than in the second step of the proof of Theorem 2.2, with an empirical

[5] In the context of UCB-V it is interesting to see the influence of the range of the rewards. More precisely, if all rewards $g_{i,t}$ are in $[0,b]$ and if one uses the upper confidence bound sequence $B_{i,s,t} = \widehat{\mu}_{i,s} + \sqrt{\frac{2\alpha V_{i,s}\log(t)}{s}} + 3b\alpha\frac{\log(t)}{s}$, then one can easily prove that the leading constant in the bound becomes $\frac{\sigma_i^2}{\Delta_i} + 2b$.

Bernstein bound (see Theorem 10.3) instead of Hoeffding-Azuma, we obtain for any $\beta \in (0,1)$:

$$\mathbb{P}(\text{(2.20) or (2.21) is true}) \leqslant \left(\frac{\log t}{\log 1/\beta} + 1\right) \frac{6}{t^{\beta\alpha}}.$$

Now we want to prove that

$$\mathbb{P}(\text{(2.22) is true and (2.23) is false}) \leqslant \frac{1}{n^{\alpha}}.$$

First note that

$$\sum_{s=1}^{t} (X_{i,s} - \mu_i)^2 - (X_{i,s} - \widehat{\mu}_{i,t})^2 = \sum_{s=1}^{t} (\widehat{\mu}_{i,t} - \mu_i)(2X_{i,t} - \mu_i - \widehat{\mu}_{i,t}) = (\mu_i - \widehat{\mu}_{i,t})^2 \geqslant 0.$$

Thus we have

$$\mathbb{P}(V_{i,T_i(t-1)} > \sigma_i^2 + \Delta_i/2) \leqslant \mathbb{P}\left(\sum_{s=1}^{T_i(t-1)} (X_{i,s} - \mu_i)^2 > \sigma_i^2 + \Delta_i/2\right).$$

Moreover note that $\mathrm{Var}((X_{i,t} - \mu_i)^2) \leqslant \sigma_i^2$. Hence, using a maximal Bernstein's inequality (see Theorem 10.2 and the remark at the beginning of the second step in the proof of Theorem 2.2), we obtain, with the notation $u = \frac{8\alpha(\sigma_i^2 + 2\Delta_i)\log n}{\Delta_i^2}$,

$$\begin{aligned}
\mathbb{P}(\text{(2.22) is true and (2.23) is false}) &\leqslant \mathbb{P}\left(\exists \ell \in \{u, \dots, t\} : \sum_{s=1}^{\ell} (X_{i,s} - \mu_i)^2 > \sigma_i^2 + \Delta_i/2\right) \\
&\leqslant \exp\left(-\frac{u(\Delta_i/2)^2}{2\sigma_i^2 + \Delta_i/3}\right) \\
&\leqslant \frac{1}{n^{\alpha}}.
\end{aligned}$$

Third step.

Mimicking step three of the proof of Theorem 2.2, we obtain (with $\beta = \frac{2}{\alpha+1}$)

$$\begin{aligned}
\mathbb{E}T_i(n) &\leqslant u + \sum_{t=u+1}^{n} \mathbb{P}(\text{(2.20) or (2.21) is true}) + \mathbb{P}(\text{(2.22) is true and (2.23) is false}) \\
&\leqslant u + \frac{1}{n^{\alpha-1}} + \sum_{t=1}^{n} \left(\frac{\log t}{\log 1/\beta} + 1\right) \frac{6}{t^{\beta\alpha}} \\
&\leqslant u + 1 + \frac{12}{(\beta\alpha - 1)^2 \log(1/\beta)},
\end{aligned}$$

which ends the proof.

□

2.3. High probability and expected regret bounds. High probability bounds are interesting for both stochastic and adversarial bandits. However for the former there exists only partial results, see Audibert et al. [2009]. In this thesis we only deal with the latter. In particular in this section we show how one can obtain a forecaster for the adversarial bandit game with high probability guarantee on its regret.

Exp3.P:

Parameters: $\eta \in \mathbb{R}^+, \gamma, \beta \in [0,1]$.

Let p_1 be the uniform distribution over $\{1, \dots, K\}$.

For each round $t = 1, 2, \dots, n$

(1) Draw an arm I_t from the probability distribution p_t.

(2) Compute the estimated loss for each arm: $\widetilde{g}_{i,t} = \frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}$ and update the estimated cumulative loss: $\widetilde{G}_{i,t} = \sum_{s=1}^{t} \widetilde{g}_{i,s}$.

(3) Compute the new probability distribution over the arms $p_{t+1} = (p_{1,t+1}, \dots, p_{K,t+1})$ where:

$$p_{i,t+1} = (1-\gamma)\frac{\exp\left(-\eta\widetilde{G}_{i,t}\right)}{\sum_{k=1}^{K}\exp\left(-\eta\widetilde{G}_{k,t}\right)} + \frac{\gamma}{K}.$$

Figure 5: Exp3.P forecaster.

Let us consider the Exp3 strategy defined in Section 2.1. This forecaster works with an estimate of the cumulative gain for each arm and this estimate is unbiased. However from the form of the estimate, one can immediately see that no interesting high probability bounds can be derived. This is obviously a trouble when one wants to get high probability bounds on the regret. One way to deal with it is to modify the estimate, and in particular to introduce a bias which permits to derive high probability statement on the estimate. More precisely, we want now an estimate of the cumulative gain which is, with high probability, an upper bound on the true cumulative gain. This goal is achieved with $\widetilde{g}_{i,t} = \frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}$ as is shown with the next lemma.

LEMMA 2.1. *Let* $\widetilde{g}_{i,t} = \frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}$ *with* $\beta \leqslant 1$. *Then with probability at least* $1-\delta$*:*

$$\sum_{t=1}^{n} g_{i,t} \leqslant \sum_{t=1}^{n} \widetilde{g}_{i,t} + \frac{\log(\delta^{-1})}{\beta}.$$

PROOF. Let $\mathbb{E}_t$ be the expectation resulting from $I_t \sim p_t$. Since $\exp(x) \leqslant 1 + x + x^2$ for $x \leqslant 1$, we have for $\beta \leqslant 1$

$$\begin{aligned}
&\mathbb{E}_t \exp\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}\right)\\
\leqslant&\left\{1 + \mathbb{E}_t\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\right) + \mathbb{E}_t\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\right)^2\right\}\exp\left(-\frac{\beta^2}{p_{i,t}}\right)\\
\leqslant&\left\{1 + \beta^2\frac{g_{i,t}^2}{p_{i,t}}\right\}\exp\left(-\frac{\beta^2}{p_{i,t}}\right)\\
\leqslant&1,
\end{aligned}$$

where the last inequality uses $1 + u \leqslant \exp(u)$. As a consequence, we have

$$\mathbb{E}\exp\left(\beta\sum_{t=1}^{n} g_{i,t} - \beta\sum_{t=1}^{n}\frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}\right) \leqslant 1.$$

Moreover Markov's inequality implies $\mathbb{P}\left(X > \log(\delta^{-1})\right) \leqslant \delta \mathbb{E} e^X$ and thus with probability at least $1-\delta$

$$\beta \sum_{t=1}^{n} g_{i,t} - \beta \sum_{t=1}^{n} \frac{g_{i,t} \mathbb{1}_{I_t = i} + \beta}{p_{i,t}} \leqslant \log(\delta^{-1}).$$

□

We describe in Figure 5 the strategy corresponding to these new estimates, which is called Exp3.P and was introduced by Auer et al. [2003]. Remark that here we directly work on the gains rather than on the modification into losses as in Section 2.1. Moreover we add a mixing with the uniform distribution [6]. Note that for sake of simplicity we also focus on the version with known time horizon, anytime results can easily be derived with the same techniques than in the proof of Theorem 2.1

We propose two different bounds. In (2.24) the algorithm needs to have the confidence level δ as a parameter. This is the usual type of bounds for Exp3.P. We slightly improve the constant with respect to Theorem 6.10 of Cesa-Bianchi and Lugosi [2006] which already improved the dependency on n with respect to the original result of Auer et al. [2003]. On the other hand (2.25) is a new type of bound, where the algorithm satisfies a high probability bound for any confidence level. This property will be particularly important to derive good bounds on the expected regret.

THEOREM 2.4 (High probability bound for Exp3.P). *Let $\delta \in (0,1)$. With $\beta = \sqrt{\frac{\log(K\delta^{-1})}{nK}}$, $\eta = 0.95\sqrt{\frac{\log(K)}{nK}}$ and $\gamma = 1.05\sqrt{\frac{K\log(K)}{n}}$, Exp3.P satisfies with probability at least $1-\delta$:*

$$R_n \leqslant 5.15\sqrt{nK\log(K\delta^{-1})}. \tag{2.24}$$

On the other hand with $\beta = \sqrt{\frac{\log(K)}{nK}}$, $\eta = 0.95\sqrt{\frac{\log(K)}{nK}}$ and $\gamma = 1.05\sqrt{\frac{K\log(K)}{n}}$, Exp3.P satisfies with probability at least $1-\delta$:

$$R_n \leqslant \sqrt{\frac{nK}{\log(K)}}\log(\delta^{-1}) + 5.15\sqrt{nK\log(K)}. \tag{2.25}$$

PROOF. We will prove that if $\gamma \leqslant 1/2$ and $(1+\beta)K\eta \leqslant \gamma$ then Exp3.P satisfies with probability at least $1-\delta$:

$$R_n \leqslant \beta nK + \gamma n + (1+\beta)\eta K n + \frac{\log(K\delta^{-1})}{\beta} + \frac{\log K}{\eta}. \tag{2.26}$$

We provide a proof of (2.26) in three steps.

First step: Notations and simple equalities.

One can immediately see that $\mathbb{E}_{i\sim p_t}\widetilde{g}_{i,t} = g_{I_t,t} + \beta K$ and thus:

$$\sum_{t=1}^{n} g_{k,t} - \sum_{t=1}^{n} g_{I_t,t} = \beta nK + \sum_{t=1}^{n} g_{k,t} - \sum_{t=1}^{n} \mathbb{E}_{i\sim p_t}\widetilde{g}_{i,t}. \tag{2.27}$$

[6] The technical reason is that in terms of losses one would need a lower bound on the true cumulative loss which would then be negative in some cases, thus Exp3.P would use the exponential function on all the reals (rather than just the negative reals) where it does not enjoy a simple upper bound in term of a quadratic function. Now since we work with gains and upper bounds on the gains we use the exponential on the positive reals and we need to mix the exponential weights with the uniform distribution to ensure that in fact we use the exponential in the interval $[0,1]$.

The key step is again to consider the log-moment of $\widetilde{g}_{i,t}$. However because of the mixing we need to introduce a few more notations. Let $U = (1/K, \dots, 1/K)$ be the uniform distribution over the arms and $w_t = \frac{p_t - U}{1-\gamma}$ be the distribution induced by Exp3.P at time t without the mixing. Then we have:

$$\begin{aligned} -\mathbb{E}_{i\sim p_t}\widetilde{g}_{i,t} &= -(1-\gamma)\mathbb{E}_{i\sim w_t}\widetilde{g}_{i,t} - \gamma\mathbb{E}_{i\sim U}\widetilde{g}_{i,t} \\ &= (1-\gamma)\left\{\frac{1}{\eta}\log\mathbb{E}_{i\sim w_t}\exp\left(\eta(\widetilde{g}_{i,t} - \mathbb{E}_{k\sim w_t}\widetilde{g}_{k,t})\right) - \frac{1}{\eta}\log\mathbb{E}_{i\sim w_t}\exp\left(\eta\widetilde{g}_{i,t}\right)\right\} \\ &\quad -\gamma\mathbb{E}_{i\sim U}\widetilde{g}_{i,t}. \end{aligned} \tag{2.28}$$

Second step: Study of the first term in (2.28)**.**
We use the inequalities $\log x \leqslant x - 1$ and $\exp(x) \leqslant 1 + x + x^2$, $\forall x \leqslant 1$ as well as the fact that $\eta\widetilde{g}_{i,t} \leqslant 1$ since $(1+\beta)\eta K \leqslant \gamma$:

$$\begin{aligned} \log\mathbb{E}_{i\sim w_t}\exp\left(\eta(\widetilde{g}_{i,t} - \mathbb{E}_{k\sim p_t}\widetilde{g}_{k,t})\right) &= \log\mathbb{E}_{i\sim w_t}\exp\left(\eta\widetilde{g}_{i,t}\right) - \eta\mathbb{E}_{k\sim p_t}\widetilde{g}_{k,t} \\ &\leqslant \mathbb{E}_{i\sim w_t}\left(\exp\left(\eta\widetilde{g}_{i,t}\right) - 1 - \eta\widetilde{g}_{i,t}\right) \\ &\leqslant \mathbb{E}_{i\sim w_t}\eta^2\widetilde{g}_{i,t}^2 \\ &\leqslant \frac{1+\beta}{1-\gamma}\eta^2\sum_{i=1}^{K}\widetilde{g}_{i,t}, \end{aligned} \tag{2.29}$$

where we used $\frac{w_{i,t}}{p_{i,t}} \leqslant \frac{1}{1-\gamma}$ for the last step.

Third step: Summing.

Set $\widetilde{G}_{i,0} = 0$. Remark that $w_t = (w_{1,t}, \dots, w_{K,t})$ with

$$w_{i,t} = \frac{\exp\left(-\eta\widetilde{G}_{i,t-1}\right)}{\sum_{k=1}^{K}\exp\left(-\eta\widetilde{G}_{k,t-1}\right)}. \tag{2.30}$$

Then plugging (2.29) in (2.28) and summing we obtain (with (2.30) too):

$$\begin{aligned} -\sum_{t=1}^{n}\mathbb{E}_{i\sim p_t}\widetilde{g}_{i,t} &\leqslant (1+\beta)\eta\sum_{t=1}^{n}\sum_{i=1}^{K}\widetilde{g}_{i,t} - \frac{1-\gamma}{\eta}\sum_{t=1}^{n}\log\left(\sum_{i=1}^{K}w_{i,t}\exp\left(\eta\widetilde{g}_{i,t}\right)\right) \\ &= (1+\beta)\eta\sum_{t=1}^{n}\sum_{i=1}^{K}\widetilde{g}_{i,t} - \frac{1-\gamma}{\eta}\log\left(\prod_{t=1}^{n}\frac{\sum_{i=1}^{K}\exp(\eta\widetilde{G}_{i,t})}{\sum_{i=1}^{K}\exp(\eta\widetilde{G}_{i,t-1})}\right) \\ &\leqslant (1+\beta)\eta K\max_j\widetilde{G}_{j,n} + \frac{\log K}{\eta} - \frac{1-\gamma}{\eta}\log\left(\sum_{t=1}^{n}\exp(\eta\widetilde{G}_{i,n})\right) \\ &\leqslant -(1-\gamma-(1+\beta)\eta K)\max_j\widetilde{G}_{j,n} + \frac{\log(K)}{\eta} \\ &\leqslant -(1-\gamma-(1+\beta)\eta K)\max_j\sum_{t=1}^{n}g_{j,t} + \frac{\log(K\delta^{-1})}{\beta} + \frac{\log(K)}{\eta}, \end{aligned}$$

where the last inequality comes from Lemma 2.1 and an union bound as well as the fact $\gamma - (1+\beta)\eta K \leqslant 1$ which is a consequence of $(1+\beta)\eta K \leqslant \gamma \leqslant 1/2$. Putting together this last inequality

and (2.27) we obtain:

$$R_n \leqslant \beta n K + \gamma n + (1+\beta)\eta K n + \frac{\log(K\delta^{-1})}{\beta} + \frac{\log(K)}{\eta},$$

which is the announced result.

(2.24) is then proved as follow. First, it is trivial if $n \geqslant 5.15\sqrt{nK\log(K\delta^{-1})}$ and thus in the following we assume that this is not the case. In particular it implies $\gamma \leqslant 0.21$ and $\beta \leqslant 0.1$ and thus we have $(1+\beta)\eta K \leqslant \gamma \leqslant 1/2$. Using (2.26) now directly yields the claimed bound. The same argument can be used to derive (2.25). □

We discuss now expected regret bounds. As the cautious reader may already have observed, in the oblivious case, a uniform (over all oblivious adversaries) bound on the pseudo-regret implies the same bound on the expected regret. This follows from noting that the expected regret against an oblivious adversary is smaller than the maximal pseudo-regret against deterministic adversaries. In the stochastic case, the following proposition allow to generalize theorems 2.2 and 2.3 to the expected regret.

PROPOSITION 2.1. *For a given $\delta \geqslant 0$, let $I = \{i \in \{1,\dots,K\} : \Delta_i \leqslant \delta\}$ be the set of arms "δ-close" to the optimal ones, and $J = \{1,\dots,K\} \setminus I$ the remaining set of arms. In the stochastic bandit game, we have*

$$\mathbb{E}R_n - \overline{R}_n \leqslant \sqrt{\frac{n\log|I|}{2}} + \sum_{i\in J}\frac{1}{2\Delta_i}\exp(-n\Delta_i^2).$$

In particular when there exists a unique arm i^* such that $\Delta_{i^*} = 0$, we have

$$\mathbb{E}R_n - \overline{R}_n \leqslant \sum_{i\neq i^*}\frac{1}{2\Delta_i}.$$

PROOF. Let $W_n^{(1)} = \max_{i\in I}\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{i^*,t}$ and $W_n^{(2)} = \max_{i\in\{1,\dots,K\}}\sum_{t=1}^n g_{i,t} - \max_{i\in I}\sum_{t=1}^n g_{i,t}$. We have $\mathbb{E}R_n - \overline{R}_n = \mathbb{E}W_n^{(1)} + \mathbb{E}W_n^{(2)}$. First we prove that

$$\mathbb{E}W_n^{(1)} \leqslant \sqrt{\frac{n\log|I|}{2}}. \tag{2.31}$$

Let $\lambda > 0$, then by Jensen's and Hoeffding's inequalities, we have:

$$\begin{aligned}
\mathbb{E}\max_{i\in I}\sum_{t=1}^n g_{i,t} &\leqslant \mathbb{E}\frac{1}{\lambda}\log\sum_{i\in I}\exp\left(\lambda\sum_{t=1}^n g_{i,t}\right)\\
&\leqslant \frac{1}{\lambda}\log\sum_{i\in I}\mathbb{E}\prod_{t=1}^n\exp(\lambda g_{i,t})\\
&= \frac{1}{\lambda}\log\sum_{i\in I}\prod_{t=1}^n\mathbb{E}\exp(\lambda g_{i,t})\\
&\leqslant \frac{1}{\lambda}\log\sum_{i\in I}\prod_{t=1}^n\exp(\lambda\mathbb{E}g_{i,t})\exp(\lambda^2/8)\\
&\leqslant \frac{\log K}{\lambda} + \max_{i\in I}\mathbb{E}\sum_{t=1}^n g_{i,t} + \frac{\lambda n}{8}.
\end{aligned}$$

Taking $\lambda = 2\sqrt{\frac{2\log K}{n}}$ ends the proof of (2.31).

Now, thanks to Hoeffding's inequality (see Theorem 10.1) and by using $\int_x^{+\infty} \exp(-u^2)du \leqslant \frac{1}{2x}\exp(-x^2)$ for any $x > 0$, we have

$$\begin{aligned}
\mathbb{E}W_n^{(2)} &= \int_0^{+\infty} \mathbb{P}(W_n^{(2)} > t)dt \\
&\leqslant \sum_{i\in J} \int_0^{+\infty} \mathbb{P}\left(\sum_{t=1}^n g_{i,t} - \max_{j\in I}\sum_{t=1}^n g_{j,t} > t\right) dt \\
&\leqslant \sum_{i\in J} \int_0^{+\infty} \mathbb{P}\left(\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{i^*,t} > t\right) dt \\
&\leqslant \sum_{i\in J} \int_0^{+\infty} \exp\left(-\frac{(t+n\Delta_i)^2}{n}\right) dt \\
&\leqslant \sum_{i\in J} \frac{1}{2\Delta_i}\exp(-n\Delta_i^2),
\end{aligned}$$

which concludes the proof. □

In the case of a non-oblivious adversary the gain vector g_t at time t is dependent of the past actions of the forecaster. This makes the proofs more intricate and to get bounds on the expected regret we first prove high probability bounds. Following the method proposed in Audibert and Bubeck [2009b], see Chapter 3, we derive a bound on the expected regret of Exp3.P using (2.25).

THEOREM 2.5 (Expected regret of Exp3.P). *With* $\beta = \sqrt{\frac{\log(K)}{nK}}$, $\eta = 0.95\sqrt{\frac{\log(K)}{nK}}$ *and* $\gamma = 1.05\sqrt{\frac{K\log(K)}{n}}$, *Exp3.P satisfies*

$$\mathbb{E}R_n \leqslant 5.15\sqrt{nK\log(K)} + \sqrt{\frac{nK}{\log(K)}}. \tag{2.32}$$

PROOF. One needs to integrate the deviations in (2.25) using the formula $\mathbb{E}W \leqslant \int_0^1 \frac{1}{\delta}\mathbb{P}(W > \log(\delta^{-1}))d\delta$ for W a real-valued random variable. In particular here taking $W = \sqrt{\frac{\log(K)}{nK}}(R_n - 5.15\sqrt{nK\log(K)})$ yields $\mathbb{E}W \leqslant 1$ which is equivalent to (2.32). □

3. Lower Bounds

It is natural to ask whether the guarantees that we have obtained in Section 2 are optimal. We propose here two different lower bounds on the pseudo-regret of any forecaster. Theorem 2.6 below, essentialy extracted from Auer et al. [2003], shows that up to a logarithmic factor the bounds (2.4), (2.5) and (2.13) are minimax optimal. In Audibert and Bubeck [2009a] the authors bridge this long open logarithmic gap between upper and lower bounds and propose a new forecaster which enjoys a $\sqrt{nK}$ pseudo-regret, matching the lower bound of Theorem 2.6, see Chapter 3 for more details. On the other hand, Theorem 2.7, extracted from Lai and Robbins [1985], shows that the bounds (2.13) and (2.19) are essentially optimal up to a constant multiplicative factor.

Both proofs are based on information theoretic tools, more precisely the Kullback-Leibler divergence, see Appendix A, Section 2 for definitions and properties.

THEOREM 2.6 (Minimax lower bound). *Let* $\sup$ *represents the supremum taken over all stochastic bandits and* $\inf$ *the infimum taken over all forecasters, then the following holds true:*

$$\inf \sup \overline{R}_n \geqslant \frac{1}{20}\sqrt{nK}, \tag{2.33}$$

and

$$\sup_{n,K} \frac{\inf \sup \overline{R}_n}{\sqrt{nK}} \geqslant \frac{1}{4}. \tag{2.34}$$

We present here a slightly simplified proof with respect to the version in Cesa-Bianchi and Lugosi [2006]. The main ideas remain the same but we introduce the empirical distribution of plays over the arm which compress the useful information of all rounds into one quantity. By doing carefully the last optimization step in the proof we also get the new bound (2.34) which improves the constant in the asymptotic regime. Since the exact dependency on n and K is known, Audibert and Bubeck [2009a], the next step is to catch the best asymptotical constant for $\inf \sup \overline{R}_n/\sqrt{nK}$. Thus (2.34) may be of interest for future work.

The general proof idea goes as follow. Since at least one arm is pulled less than n/K times, for this arm one can not differentiate between a Bernoulli of parameter $1/2$ and $1/2 + \sqrt{K/n}$. Thus if all arms are Bernoulli of parameter $1/2$ but one with parameter $1/2 + \sqrt{K/n}$, then the forecaster should incur a regret of order $n\sqrt{K/n} = \sqrt{nK}$. To formalize this idea we use the Kullback-Leibler divergence, and in particular Pinsker's inequality to compare the behavior of a given forecaster on the null bandit (where all arms are Bernoulli of parameter $1/2$) and the same bandit where we raise by ε the parameter of one arm.

We prove a lemma which will be useful on its own to derive lower bounds in other contexts. The proof of Theorem 2.6 then follows by a careful optimization over ε.

LEMMA 2.2. *Let* $\varepsilon \in [0, 1)$. *For any* $i \in \{1, \dots, K\}$ *let* $\mathbb{E}_i$ *be the expectation under the bandit with all arms being Bernoulli of parameter* $\frac{1-\varepsilon}{2}$ *but arm* i *has parameter* $\frac{1+\varepsilon}{2}$. *Then for any forecaster the following holds true:*

$$\sup_{i=1,\dots,K} \mathbb{E}_i \sum_{t=1}^{n} (g_{i,t} - g_{I_t,t}) \geqslant n\varepsilon \left(1 - \frac{1}{K} - \sqrt{\varepsilon \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)}\sqrt{\frac{n}{2K}}\right).$$

PROOF. We provide a proof in five steps by lower bounding $\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_i \sum_{t=1}^{n} (g_{i,t} - g_{I_t,t})$. This will imply the statement of the lemma since a sup is larger than a mean.

First step: Empirical distribution of plays.

Until the fifth step we consider a deterministic forecaster, that is he does not have access to an external randomization. Let $q_n = (q_{1,n}, \dots, q_{K,n})$ be the empirical distribution of plays over the arms defined by:

$$q_{i,n} = \frac{T_i(n)}{n}.$$

Let J_n be drawn according to q_n. We note $\mathbb{P}_i$ the law of J_n under the bandit with all arms being Bernoulli of parameter $\frac{1-\varepsilon}{2}$ but arm i has parameter $\frac{1+\varepsilon}{2}$ (we call it the i-th bandit). Remark that

we have $\mathbb{P}_i(J_n = j) = \mathbb{E}_i T_j(n)/n$, hence,

$$\mathbb{E}_i \sum_{t=1}^{n} (g_{i,t} - g_{I_t,t}) = \varepsilon n \sum_{j \neq i} \mathbb{P}_i(J_n = j) = \varepsilon n (1 - \mathbb{P}_i(J_n = i))$$

which implies

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_i \sum_{t=1}^{n} (g_{i,t} - g_{I_t,t}) = \varepsilon n \left(1 - \frac{1}{K} \sum_{i=1}^{K} \mathbb{P}_i(J_n = i) \right). \tag{2.35}$$

Second step: Pinsker's inequality.

Let $\mathbb{P}_0$ be the law of J_n under the bandit with all arms being Bernoulli of parameter $\frac{1-\varepsilon}{2}$ (we call it the 0-th bandit). Then Lemma 10.2 directly gives:

$$\mathbb{P}_i(J_n = i) \leqslant \mathbb{P}_0(J_n = i) + \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)}.$$

Hence

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{P}_i(J_n = i) \leqslant \frac{1}{K} + \frac{1}{K} \sum_{i=1}^{K} \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)}. \tag{2.36}$$

Third step: Computation of $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)$ with the Chain rule for Kullback-Leibler divergence. Remark that since the forecaster is deterministic, the sequence of rewards $W_n = (Y_1, \ldots, Y_n) \in \{0,1\}^n$ received by the algorithm uniquely determines the empirical distribution of plays q_n, in particular the law of J_n conditionnaly to W_n is the same for any bandit. Thus if for $i \in \{0, \ldots, K\}$ we note $\mathbb{P}_i^n$ the law of W_n under the i-th bandit then one can easily prove, with Lemma 10.4 for instance, that

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i) \leqslant \mathrm{KL}(\mathbb{P}_0^n, \mathbb{P}_i^n). \tag{2.37}$$

Now we use Lemma 10.4 iteratively to introduce the laws $\mathbb{P}_i^t$ of $W_t = (Y_1, \ldots, Y_t)$. More precisely we have:

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_0^n, \mathbb{P}_i^n) \\
&= \mathrm{KL}(\mathbb{P}_0^1, \mathbb{P}_i^1) + \sum_{t=2}^{n} \sum_{w_{t-1}} \mathbb{P}_0^{t-1}(w_{t-1}) \mathrm{KL}(\mathbb{P}_0^t(.|w_{t-1}), \mathbb{P}_i^t(.|w_{t-1})) \\
&= \mathrm{KL}(\mathbb{P}_0^1, \mathbb{P}_i^1) + \sum_{t=2}^{n} \left\{ \sum_{w_{t-1}: I_t = i} \mathbb{P}_0^{t-1}(w_{t-1}) \mathrm{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2} \right) \right. \\
&\qquad \left. + \sum_{w_{t-1}: I_t \neq i} \mathbb{P}_0^{t-1}(w_{t-1}) \mathrm{KL}\left(\frac{1+\varepsilon}{2}, \frac{1+\varepsilon}{2} \right) \right\} \\
&= \mathrm{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2} \right) \mathbb{E}_0 T_i(n). \qquad (2.38)
\end{aligned}$$

Fourth step: Conclusion for deterministic forecasters with the concavity of the square root. By using that the square root is concave and combining (2.37) and (2.38) we deduce:

$$\begin{aligned}\frac{1}{K}\sum_{i=1}^{K}\sqrt{\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_i)} &\leqslant \sqrt{\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_i)} \\ &\leqslant \sqrt{\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)\mathbb{E}_0 T_i(n)} \\ &= \sqrt{\frac{n}{K}\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)}. \qquad (2.39)\end{aligned}$$

To conclude the proof for deterministic forecaster one needs to plug in (2.36) and (2.39) in (2.35) along with the following simple computations:

$$\begin{aligned}\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right) &= (1-\varepsilon)/2\log\left(\frac{1-\varepsilon}{1+\varepsilon}\right)+(1+\varepsilon)/2\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \\ &= \varepsilon\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right).\end{aligned}$$

Fifth step: Fubini's Theorem to handle non-deterministic forecasters.

Now let us consider a randomized forecaster. Denote by $\mathbb{E}_{reward,i}$ the expectation with respect to the reward generation process of the i-th bandit, $\mathbb{E}_{rand}$ the expectation with respect to the randomization of the strategy and $\mathbb{E}_i$ the expectation with respect to both processes. Then one has (thanks to Fubini's Theorem)

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t}) = \mathbb{E}_{rand}\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{reward,i}\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t}).$$

Now remark that if we fix the realization of the forecaster's randomization then the results of the previous steps apply and in particular we can lower bound $\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{reward,i}\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t})$ as before. □

The next theorem gives a distribution-dependent lower bound for the stochastic bandit game. We provide a statement less general than the original version of Lai and Robbins [1985] but it suits our purposes. The proof follows the exact same lines than the original. Moreover Lemma 10.3 enables us to compare Theorem 2.7 with (2.13) and (2.19).

THEOREM 2.7 (Distribution-dependent lower bound in the stochastic bandit game). *Let us consider a forecaster such that for any stochastic bandit, any arm i such that $\Delta_i > 0$ and any $a > 0$, we have $\mathbb{E}T_i(n) = o(n^a)$. Then for any stochastic bandit with Bernoulli distributions, all different from a Dirac distribution at 1, the following holds true:*

$$\liminf_{n\to+\infty}\frac{\overline{R}_n}{\log n} \geqslant \sum_{i:\Delta_i>0}\frac{\Delta_i}{\mathrm{KL}(\mu_i,\mu^*)}.$$

PROOF. We provide a proof in three steps.

First step: Notations.
Without loss of generality let us assume that arm 1 is optimal and arm 2 is suboptimal, that is $\mu_2 < \mu_1 < 1$. Let $\varepsilon > 0$. Since $x \mapsto \mathrm{KL}(\mu_2, x)$ is continuous one can find $\mu_2' \in (\mu_1, 1)$ such that

$$\mathrm{KL}(\mu_2, \mu_2') \leqslant (1+\varepsilon)\mathrm{KL}(\mu_2, \mu_1). \tag{2.40}$$

We note $\mathbb{E}', \mathbb{P}'$ when we integrate with respect to the modified bandit where the parameter of arm 2 is replaced by μ_2'. We want to compare the behavior of the forecaster on the initial and modified bandits. In particular we prove that with a fair probability the forecaster can not distinguish between the two problems. Then using the fact that we have a good forecaster (by hypothesis in the Theorem) we know that the algorithm does not make too much mistakes on the modified bandit where arm 2 is optimal, in other words we have a lower bound on the number of times the optimal arm is played. This reasoning implies a lower bound on the number of times arm 2 is played in the initial problem.

To complete this program we introduce a few notations. Recall that $X_{2,1}, \ldots, X_{2,n}$ is the sequence of random variables obtained while pulling arm 2. For $s \in \{1, \ldots, n\}$, let

$$\widehat{\mathrm{KL}}_s = \sum_{t=1}^{s} \log\left(\frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right).$$

In particular note that with respect to the initial bandit, $\widehat{\mathrm{KL}}_{T_2(n)}$ is the (non re-normalized) empirical estimation of $\mathrm{KL}(\mu_2, \mu_2')$ at time n since in that case (X_s) is i.i.d from a Bernoulli of parameter μ_2. An other important property is that for any event A one has:

$$\mathbb{P}'(A) = \mathbb{E}\, \mathbb{1}_A \exp\left(-\widehat{\mathrm{KL}}_{T_2(n)}\right). \tag{2.41}$$

Now to control the link between the behavior of the forecaster on the initial and modified bandits we introduce the event:

$$C_n = \left\{T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')}\log(n) \text{ and } \widehat{\mathrm{KL}}_{T_2(n)} \leqslant (1-\varepsilon/2)\log(n)\right\}. \tag{2.42}$$

Second step: $\mathbb{P}(C_n) = o(1)$.

By (2.41) and (2.42) one has:

$$\mathbb{P}'(C_n) = \mathbb{E}\, \mathbb{1}_{C_n} \exp\left(-\widehat{\mathrm{KL}}_{T_2(n)}\right) \geqslant \exp\left(-(1-\varepsilon/2)\log(n)\right)\mathbb{P}(C_n),$$

which implies by (2.42) and Markov's inequality:

$$\mathbb{P}(C_n) \leqslant n^{(1-\varepsilon/2)}\mathbb{P}'(C_n) \leqslant n^{(1-\varepsilon/2)}\mathbb{P}'\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')}\log(n)\right) \leqslant n^{(1-\varepsilon/2)}\frac{\mathbb{E}'(n - T_2(n))}{n - \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')}\log(n)}.$$

Now remark that in the modified bandit arm 2 is the unique optimal arm, thus our assumption that for any bandit, any suboptimal arm i, any $a > 0$, one has $\mathbb{E}T_i(n) = o(n^a)$ implies that

$$\mathbb{P}(C_n) \leqslant n^{(1-\varepsilon/2)}\frac{\mathbb{E}'(n - T_2(n))}{n - \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')}\log(n)} = o(1).$$

Third step: $\mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')}\log(n)\right) = o(1)$.

Remark that

$$\begin{aligned}
&\mathbb{P}(C_n) \\
&\geqslant \mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')}\log(n) \text{ and } \max_{1\leqslant s\leqslant \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')}\log(n)} \widehat{\mathrm{KL}}_s \leqslant (1-\varepsilon/2)\log(n)\right) \\
&= \mathbb{P}\Bigg(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')}\log(n) \\
&\qquad \text{and } \frac{\mathrm{KL}(\mu_2,\mu_2')}{(1-\varepsilon)\log(n)} \max_{1\leqslant s\leqslant \frac{(1-\varepsilon)\log(n)}{\mathrm{KL}(\mu_2,\mu_2')}} \widehat{\mathrm{KL}}_s \leqslant \frac{1-\varepsilon/2}{1-\varepsilon}\mathrm{KL}(\mu_2,\mu_2')\Bigg). \qquad (2.43)
\end{aligned}$$

Now using Lemma 10.5 since $\mathrm{KL}(\mu_2,\mu_2') > 0$ and the fact that $\frac{1-\varepsilon/2}{1-\varepsilon} > 1$ we deduce that

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\mathrm{KL}(\mu_2,\mu_2')}{(1-\varepsilon)\log(n)} \max_{1\leqslant s\leqslant \frac{(1-\varepsilon)\log(n)}{\mathrm{KL}(\mu_2,\mu_2')}} \widehat{\mathrm{KL}}_s \leqslant \frac{1-\varepsilon/2}{1-\varepsilon}\mathrm{KL}(\mu_2,\mu_2')\right) = 1,$$

and thus by the result of the second step and (2.43):

$$\mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')}\log(n)\right) = o(1).$$

Now using (2.40) we obtain:

$$\mathbb{E}T_2(n) \geqslant (1+o(1))\frac{1-\varepsilon}{1+\varepsilon}\frac{\log(n)}{\mathrm{KL}(\mu_2,\mu_1)}$$

which concludes the proof by (2.12). □

4. Extensions

In this section, we discuss some basic extensions of the multi-armed bandit problem, both in stochastic and adversarial settings. Some of the concrete applications that motivate these extensions were presented in Section 1.3. An exhaustive list of extensions is beyond the scope of this introduction. We only cite papers introducing each specific extension (in the bandit context) and papers with state of the art results.

4.1. Large set of arms. In this section we consider works which investigate the consequences of enlarging the set of arms. At each time step the forecaster chooses an arm in a (possibly continuous) set $\mathcal{X}$. We still evaluate the performance of a forecaster through its regret with respect to the best single arm. In the multi-armed bandit problem, one has $\mathcal{X} = \{1,\dots,K\}$ and the number of rounds is larger than the number of arms. Here we want to consider much larger sets $\mathcal{X}$ which enjoy some internal structure. More precisely, in the stochastic version we assume some regularity on the mean reward function (defined over $\mathcal{X}$) and for the adversarial version we constraint the choice of the rewards for the adversary.

Stochastic setting.

- **$\mathcal{X}$-Armed Bandits**, Bubeck et al. [2009c]. To have a gentle start let us consider a more restrictive framework than the one described in Chapter 4 but which shall be enough to convey the idea of an $\mathcal{X}$-armed bandit. Let us assume that $\mathcal{X}$ is a metric space and that the mean reward function f (which maps arms to the average gain one receives by pulling this arm) is 1-Lipschitz. In Kleinberg et al. [2008a] the authors defines a notion

of intrinsic dimension d of f (which we call near-optimality dimension) and prove that one can obtain a pseudo-regret of order $n^{\frac{d+1}{d+2}}$. In Chapter 4 we extend this result to more general spaces than metric and we weaken the Lipschitz assumption. Moreover we exhibit the minimax rate for the regret and show that our algorithm attains it.

- **Many Armed Bandits**, Berry et al. [1997]. In this context the forecaster is facing a countably infinite number of arms with Bernoulli's distributions. The usual assumption in this context is that the forecaster has a prior on the parameters of the arms. However this assumption was recently weakened in Wang et al. [2009], where the authors assume that each arm has a probability of order ε^β to be ε optimal. The authors then derive regret bounds which depend on the parameter β.

Adversarial setting.

- **Combinatorial Bandits**, Cesa-Bianchi and Lugosi [2009]. In this setting the set of arms is a subset of the hypercube, $\mathcal{X} \subset \{0,1\}^d$, and the adversary chooses a vector $\bar{g}_t \in [0,1]^d$. The gain of an arm $x \in \mathcal{X}$ is then defined as the inner product $\bar{g}_t^T x$. Let $|\mathcal{X}| = K$ and $B = \max_{x \in \mathcal{X}} ||x||_1$. The authors propose an algorithm achieving in most cases (see Theorem 1, Cesa-Bianchi and Lugosi [2009], for the details) a pseudo-regret of order $B\sqrt{nd\log(K)}$. Remark that one could use naively the Exp3 strategy to get a pseudo-regret of order $B\sqrt{nK\log(K)}$. However we would like to handle very large K (typically K is exponential in d). Thus the former bound is much more interesting. An example of particular importance which fits in this framework is the online shortest path problem described in Section 1.3.
- **Bandit Linear Optimization**, Awerbuch and Kleinberg [2004]. This problem can be viewed as a generalization of Combinatorial Bandits (though it was considered first). Here the set of arms is a compact and convex set $\mathcal{X} \subset \mathbb{R}^d$, and the adversary chooses $\bar{g}_t \in \mathbb{R}^d$ so that the reward of arm $x \in \mathcal{X}$ is $\bar{g}_t^T x \in [-1,1]$. In Abernethy et al. [2008] the authors propose a computationally efficient algorithm achieving a pseudo-regret of order $poly(d)\sqrt{n\log n}$. In the particular case of $\mathcal{X}$ being the unit euclidean ball the bound is actually $d\sqrt{n\log n}$, and a matching lower bound (up to a logarithmic term in n) is proposed in Dani et al. [2008]. Note that this problem has also been considered in a stochastic setting, see *e.g.*, Dani et al. [2008], Abbasi-Yadkori [2009] and Rusmevichientong and Tsitsiklis [2009].

4.2. Different notions of regret. The behavior of the forecaster in the classical multi-armed bandit problem is guided by its goal of minimizing the cumulative regret. By defining different notions of regret one can dramaticaly change the type of optimal behavior for the forecaster. We present here a few other notions of regret which has been recently studied.

Stochastic setting.

- **Simple Regret**, Bubeck et al. [2009a]. In Chapter 6 we introduce a new notion of regret for the classical multi-armed bandit problem, called simple regret. We ask the forecaster to output an arm J_n based on the information collected during the n rounds of plays of a classical stochastic bandit game. We evaluate the performance of the forecaster only through the arm J_n, more precisely we consider the simple regret $r_n = \mu^* - \mu_{J_n}$. That is, during the n rounds of plays the forecaster is not constrained by the goal of maximizing his cumulative gain and can explore "freely" the set of arms to find the best one. We argue in chapters 6 and 7 that this kind of regret is very relevant in many applications of

bandits algorithms and we propose an almost complete analysis both in terms of upper and lower bounds.

- **Active Learning Regret**, Antos et al. [2008]. Here one models a situation where we want to estimate well the mean of all arms. More precisely the forecaster aims at minimizing the objective function $\max_{i\in\{1,\dots,K\}} \mathbb{E}(\mu_i - \widehat{\mu}_{i,T_i(n)})^2$ with the notations of Section 2.2. The regret on a particular bandit problem is then defined as the difference between the best objective value one can obtain with a static allocation strategy (that is $T_i(n)$ is not a random variable but is fixed beforehand) and the actual value achieved by the forecaster. The authors propose an algorithm achieving an expected regret of order $n^{-3/2}$ and conjecture that this rate is optimal.
- **Max Regret**, Streeter and Smith [2006a]. In this setting one aims at maximizing the maximal reward obtained at the end of the n rounds. The regret is then defined as the difference between the maximal reward one could have obtained by pulling only the best single arm and the maximal reward actually obtained. In Streeter and Smith [2006b] the authors propose a forecaster with a vanishing expected regret as the number of rounds tends to infinity.

Adversarial setting.

- **Tracking the best expert regret**, Auer [2002]. In the traditional cumulative regret we compare the forecaster to the best fixed arm in hindsight. For some cases, bounds on this regret are rather weak statement and we would like to be able to track the best arm at different rounds. The tracking the best expert regret compares the gain of our strategy with the gain of the best switching strategy allowed to switch S times between different arms. For $S = 0$ we recover the classical regret, and for $S > 0$ we get a stronger notion of regret. The best known bound was of order $\sqrt{nKS\log(nK)}$, Auer [2002], and in Chapter 3 we improve it to $\sqrt{nKS\log(\frac{nK}{S})}$.
- **Expert Regret**, Auer et al. [2003]. Here we assume that, in addition to a classical multi-armed bandit problem we have a set of N experts which give an advice on the arm to be played at each round. The expert regret is then defined with respect to the best single expert in hindsight. In Auer et al. [2003] the authors propose an algorithm, Exp4, which achieves an expert regret of order $\sqrt{nK\log N}$, whereas using naively the Exp3 strategy on the set of experts would have a regret of order $\sqrt{nN\log N}$. Thus, in this framework we can be competitive even with an exponentially large (with respect to the number of rounds) number of experts.
- **Internal Regret**, Foster and Vohra [1997]. Here we study the regret of not having played arm j instead of arm i. More precisely the internal regret is the maximum such regret over the pair (i, j). This notion of regret is particularly important in the context of Game Theory, because of its link with the convergence to correlated equilibria, see Cesa-Bianchi and Lugosi [2006]. In Stoltz [2005] the author proposes an algorithm achieving an internal regret of order $\sqrt{nK\log K}$.

4.3. Additional rules. While the multi-armed bandit problem could potentially be applied in a variety of domains, it suffers from lack of flexibility. For most real life applications a bandit problem comes with some additional constraints or some prior knowledge. In this section we present works which investige bandits problems with these additional features.

Stochastic setting.

- **Restless Bandits**, Whittle [1988]. In its general formulation this problem considers a stochastic multi-armed bandit where the distributions on the arms are changing over time. By constraining the evolution process of the distributions one can compete with the strategy playing the best arm at each single time step (on the contrary to the adversarial bandit where we only compare to the best fixed arm). Recent works for this problem include Garivier and Moulines [2008], Slivkins and Upfal [2008].
- **Sleeping Experts**, Kleinberg et al. [2008b], **Mortal Bandits**, Chakrabarti et al. [2009]. Theses works consider the case where the set of arms is varying over time. In the former there exists a fixed set of arms but they are not all available at all rounds, the regret is then defined with respect to a strategy pulling the best available arm at each time step. In the latter each arm has a given budget and is replaced by a new arm once this budget is exhausted. Assuming a prior distribution on the arm generation process one can define the maximal average gain per round that any strategy can obtain in the asymptotic regime, which in turns allow to define a notion of regret.

Adversarial setting.

- **Full Information (FI) Game**, Cesa-Bianchi et al. [1997]. Historically the adversarial bandit is an extension of the full information game. Here at the end of each round the adversary reveals the full gain vector g_t rather than only the gain of the arm choosed by the forecaster in the bandit game. In this setting the best expected regret one can obtain is of order $\sqrt{n \log K}$, see Chapter 3 for details.
- **Label Efficient (LE) Game**, Cesa-Bianchi et al. [2005], Allenberg et al. [2006]. This is an extension for both the full information (FI) game and the adversarial bandit. At each round the forecaster can ask to see the reward(s) with the global constraint that he can not ask it more than m times over the n rounds. In Chapter 3 we improve the expected regret upper bound for the LE-FI game to $n\sqrt{\frac{\log K}{m}}$ (it was previously known for the pseudo-regret, Cesa-Bianchi et al. [2005]) as well as the pseudo-regret bound for the LE bandit game to $n\sqrt{\frac{K}{m}}$, see Chapter 3 for details.
- **Side Information**, Wang et al. [2005]. In this setting at each round we are given an additional information $x \in \mathcal{X}$. The game goes as usually, but we compare the forecaster to the best fixed hypothesis $h : \mathcal{X} \to \{1, \ldots, K\}$ in a class $\mathcal{H}$. In Ben-David et al. [2009] the authors show that one can achieve a regret bound of order $\sqrt{n\mathrm{Ldim}(\mathcal{H})}$ where $\mathrm{Ldim}(\mathcal{H})$ is the Littlestone dimension of $\mathcal{H}$. In Lazaric and Munos [2009] it is proved that if one makes a stochastic assumption on the side information (rather than adversarial as in Ben-David et al. [2009]), then one can replace the Littlestone dimension by the usual Vapnik-Chervonenkis dimension (which can be significantly smaller).

4.4. Planning, Reinforcement Learning. In most cases, by combining some of the previous extensions, we obtain a Reinforcement Learning (RL) problem, see *e.g.*, Sutton and Barto [1998] and Kakade [2003]. The most commonly used model in RL is the Markov Decision Process (MDP), which can be seen as a stochastic restless bandit with side information where the evolution of the distributions is determined by our actions. More precisely, at each time step the forecaster receives a state information x_t and chooses an action a_t among the set $\{1, \ldots, K\}$. The forecaster then receives a reward r_t drawn from a probability distribution depending on both a_t and x_t and moves to a next state x_{t+1} (which is also generated from a probability distribution depending on both a_t and x_t). The forecaster's goal is to maximize his discounted sum of rewards $\sum_{t=1}^{+\infty} \gamma^t r_t$ where $\gamma \in (0, 1)$ is a discount factor. One can then compute his regret with respect to an optimal

strategy (which is a map between states and actions with highest expected discounted sum of rewards from each state). Obviously in this setting one has to give some flexibility to the forecaster to have a chance to find the optimal policy. The most simple assumption is that he can play several sequences of actions, each time re-starting at the initial state x_1.

In Chapter 5, we propose a new approach to this problem. We consider a weaker notion of regret than the one which compares to the optimal policy. In control terminology, we compare ourselves to the optimal open-loop policy (a mapping from time to actions rather than from states to actions). In this context we propose new bounds, and prove in particular the minimax optimality of our strategy OLOP (Open Loop Optimistic Planning).

CHAPTER 3

Minimax Policies for Bandits Games

This chapter deals with four classical prediction games, namely full information, bandit and label efficient (full information or bandit) games as well as four different notions of regret: pseudo-regret, expected regret, high probability regret and tracking the best expert regret. We introduce a new forecaster, INF (Implicitly Normalized Forecaster) based on an arbitrary function ψ for which we propose a unified analysis of its pseudo-regret in the four games we consider. In particular, for $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$, INF reduces to the classical exponentially weighted average forecaster and our analysis of the pseudo-regret recovers known results while for the expected regret we slightly tighten the bounds. On the other hand with $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$, which defines a new forecaster, we are able to remove the extraneous logarithmic factor in the pseudo-regret bounds for bandits games, and thus fill in a long open gap in the characterization of the minimax rate for the pseudo-regret in the bandit game.

We also consider the stochastic bandit game, and prove that an appropriate modification of the upper confidence bound policy UCB1 (Auer et al., 2002) achieves the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays.

Contents

This chapter is a joint work with Jean-Yves Audibert. It is based on the extended version Audibert and Bubeck [2009b] (currently under submission) of Audibert and Bubeck [2009a] which appeared in the proceedings of the 22nd Annual Conference on Learning Theory.

1. Introduction

We consider a general prediction game where at each stage, a forecaster (or decision maker) chooses one action (or arm), and receives a reward from it. Then the forecaster receives a feedback about the rewards which he can use to make his choice at the next stage. His goal is to maximize his cumulative gain. In the simplest version, after choosing an arm the forecaster observes the rewards for all arms, this is the so called full information game. Another classical example is the bandit

The prediction games:

Parameters: the number of arms (or actions) K and the number of rounds n with $n \geqslant K \geqslant 2$.

For each round $t = 1, 2, \ldots, n$

(1) The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.

(2) Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$ (see Section 6 for loss games or signed games).

(3) The forecaster receives the gain $g_{I_t,t}$ (without systematically observing it). He observes
- the reward vector $(g_{1,t}, \ldots, g_{K,t})$ in the **full information** game,
- the reward vector $(g_{1,t}, \ldots, g_{K,t})$ if he asks for it with the global constraint that he is not allowed to ask it more than m times for some fixed integer number $1 \leqslant m \leqslant n$. This prediction game is the **label efficient** game,
- only $g_{I_t,t}$ in the **bandit** game,
- only his obtained reward $g_{I_t,t}$ if he asks for it with the global constraint that he is not allowed to ask it more than m times for some fixed integer number $1 \leqslant m \leqslant n$. This prediction game is the **bandit label efficient** game.

Goal : The forecaster tries to maximize his cumulative gain $\sum_{t=1}^{n} g_{I_t,t}$.

Figure 1: The four prediction games considered in this work.

game, described in Chapter 2, where the forecaster only observes the reward of the arm he has chosen. In its original version, Robbins [1952], this game was considered in a stochastic setting, i.e., the nature draws the rewards from a fixed product-distribution. Later it was considered in the game-theoretic framework, Auer et al. [1995], where there is an adversary choosing the rewards on the arms. A classical extension of these games is the label efficient setting, Cesa-Bianchi et al. [2005] ,where you can ask for the feedback only a limited number of times. These four games are described more precisely in Figure 1.

A natural way to assess the performance of a forecaster in these games is to compute his *regret* with respect to the best action in hindsight (see Section 5 for a more general regret in which we compare to the best switching strategy having a fixed number of action-switches):

$$R_n = \max_{i=1,\ldots,K} \sum_{t=1}^{n} \big(g_{i,t} - g_{I_t,t}\big).$$

A lot of attention has been drawn by the exact characterization of the minimax expected regret in the different games we have described. More precisely for a given game, let us write $\sup$ for the supremum over all allowed adversaries and $\inf$ for the infimum over all forecaster strategies for this game. We are interested in the quantity:

$$\inf \sup \mathbb{E} R_n,$$

where the expectation is with respect to the possible randomization of the forecaster and the adversary. Another related quantity which can be easier to handle is the *pseudo-regret*:

$$\overline{R}_n = \max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} \big(g_{i,t} - g_{I_t,t}\big).$$

Despite numerous works, there were still several logarithmic gaps between upper and lower bounds on the minimax rate, namely:

(1) $\sqrt{\log(K)}$ (respectively $\sqrt{\log(n)}$) gap for the minimax pseudo-regret (respectively expected regret) in the bandit game as well as the label efficient bandit game.
(2) $\sqrt{\log(n)/\log(K)}$ gap for the minimax expected regret in the label efficient full information game,

To be more precise, we hereafter provide known results relative to these gaps. For sake of completeness we also provide the results for the full information (respectively F.I. label efficient) where there is no gap for the expected regret (respectively for the pseudo-regret). Apart from the full information game, the bounds are usually proved on the pseudo-regret. Another type of bounds which are not reproduced here are the high probability bounds on the regret. Usually in this case the parameters of the algorithm depends on the confidence level δ that we want to obtain. Thus to derive bounds on the expected regret we can not integrate the deviations but rather we have to take δ of order $1/n$, which leads to the gaps involving $\log(n)$. Note also that in the following theorems we provide the parameters used in the corresponding papers.

THEOREM 3.1. *We consider here the full information game. The exponentially weighted average forecaster with* $\eta = \sqrt{8\log K/n}$ *satisfies*

$$\sup \mathbb{E}R_n \leqslant \sqrt{(n/2)\log K}.$$

Besides we have

$$\sup_{n,K} \inf \sup \frac{\mathbb{E}R_n}{\sqrt{(n/2)\log K}} \geqslant 1.$$

The upper bound comes from the analysis in Cesa-Bianchi [1999], while the lower bound is due to Cesa-Bianchi et al. [1997].

THEOREM 3.2 (Cesa-Bianchi et al., 2005). *We consider here the label efficient full information game. The label efficient exponentially weighted average forecaster with* $\varepsilon = m/n$ *and* $\eta = \frac{\sqrt{2m\log K}}{n}$ *satisfies*

$$\sup \overline{R}_n \leqslant n\sqrt{\frac{2\log K}{m}}.$$

Besides we have for all $n \geqslant m \geqslant 14.6\log(K-1)$*:*

$$\inf \sup \overline{R}_n \geqslant 0.1\, n\sqrt{\frac{\log(K-1)}{m}}.$$[1]

THEOREM 3.3 (Auer et al., 1995). *We consider here the bandit game. The EXP3 policy with* $\eta = \sqrt{\frac{2\log K}{nK}}$ *and* $\gamma = \min\left(1, \sqrt{\frac{K\log K}{(e-1)n}}\right)$ *satisfies*

$$\sup \overline{R}_n \leqslant 2\sqrt{(e-1)nK\log K}.$$

Besides we have

$$\inf \sup \overline{R}_n \geqslant \frac{1}{20}\sqrt{nK}.$$

THEOREM 3.4 (Allenberg et al., 2006). *We consider here the label efficient bandit game. The GREEN forecaster with* $\varepsilon = m/n$, $\eta = \frac{2}{n}\sqrt{\frac{m\log K}{K}}$ *and* $\gamma = \frac{1}{K(m+2)}$ *satisfies*

$$\sup \overline{R}_n \leqslant 4n\sqrt{\frac{K\log K}{m}} + \frac{n(K\log K + 2)}{m} + \frac{n\log(m+2)}{m}.$$

[1] This bound does not give any information in the case $K = 2$. However one can modify the proof and replace Fano's lemma by the version of Birgé (2006), leading to a bound in $\log(K)$ instead of $\log(K-1)$.

	$\inf \sup \overline{R}_n$		$\inf \sup \mathbb{E}R_n$		H.P. bound on R_n
	Lower bound	Upper bound	Lower bound	Upper bound	Upper bound
Full Information	$\sqrt{n \log K}$	$\sqrt{n \log K}$	$\sqrt{n \log K}$	$\sqrt{n \log K}$	$\sqrt{n \log(K\delta^{-1})}$
Label Efficient	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$\mathbf{n\sqrt{\frac{\log K}{m}}}$	$\mathbf{n\sqrt{\frac{\log(K\delta^{-1})}{m}}}$
Oblivious Bandit[2]	$\sqrt{nK}$	$\mathbf{\sqrt{nK}}$	$\sqrt{nK}$	$\mathbf{\sqrt{nK}}$	$\mathbf{\sqrt{nK}\log(\delta^{-1})}$
General Bandit	$\sqrt{nK}$	$\mathbf{\sqrt{nK}}$	$\sqrt{nK}$	$\mathbf{\sqrt{nK \log K}}$	$\mathbf{\sqrt{\frac{nK}{\log K}}\log(K\delta^{-1})}$
L.E. Bandit		$\mathbf{n\sqrt{\frac{K}{m}}}$		$\mathbf{n\sqrt{\frac{K \log K}{m}}}$	$\mathbf{n\sqrt{\frac{K}{m \log K}}\log(K\delta^{-1})}$

Table 1: Bounds on the pseudo-regret, expected regret and high probability bounds on the regret. In bold red, the cells in which we improve the best known bound. The high probability bounds improve upon previous works because the proposed algorithms do not depend on the confidence level δ.

Contributions of this work. We propose a new forecaster, INF (Implicitely Normalized Forecaster), for which we propose a unified analysis of its pseudo-regret in the four games we consider. The analysis is original (it avoids the traditional but scope-limiting argument based on the simplification of a sum of logarithms of ratios), and leads to the following new results:

(1) We fill in the long open gap of Theorem 3.3 and prove that INF with $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ has a pseudo-regret of order $\sqrt{nK}$ (for well chosen parameters η, γ and q).

(2) With a careful analysis, we derive high probability bounds on the regret in each games with algorithms independent of the confidence level. In particular, we can do this for the well-known exponentially weighted average forecaster which corresponds in our setting to $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$. For the bandit game (respectively label efficient full information and label efficient bandit), this allows us to derive a bound of order $\sqrt{nK \log K}$ on the expected regret (respectively $n\sqrt{\frac{\log K}{m}}$ and $n\sqrt{\frac{K \log K}{m}}$) instead of the known $\sqrt{nK \log n}$ (respectively $n\sqrt{\frac{\log n}{m}}$ and $n\sqrt{\frac{K \log n}{m}}$). In the label efficient full information, this bridges the gap between the upper bound and the lower bound. We also conjecture that this is the true rate for the expected regret in the bandit and label efficient bandit games against a non-oblivous adversary[2]. Table 1 recaps the bounds for the pseudo-regret and the expected regret in the different games we consider.

(3) We prove a regret bound of order $\sqrt{nKS \log(\frac{nK}{S})}$ when we compare ourselves to a strategy allowed to switch S times between arms while the best known bound was $\sqrt{nKS \log(nK)}$, Auer [2002].

We also consider the stochastic bandit game where we prove with a modification of UCB1, Auer et al. [2002], that it is possible to attain the optimal distribution-free rate $\sqrt{nK}$ as well as the logarithmic distribution-dependent rate.

Outline.

[2]We say that an adversary is oblivious if its choices of the rewards do not depend on the past draws and obtained rewards.

(1) In Section 2 we describe a new class of forecasters for prediction games. Then we present two particular forecasters, Exp INF (which coincides with the exponentially weighted average forecaster) and Poly INF (a new forecaster), for which we propose two general theorems bounding their regret. A more general statement on the regret of any forecaster in the class we consider can be found in Section 8.
(2) In Section 3 we prove that our forecasters and analysis recover the known results for the full information game.
(3) Section 4 contains the main contributions of this chapter, namely all the regret bounds for the limited feedback games.
(4) In Section 5 we consider a stronger notion of regret, when we compare ourselves to a strategy allowed to switch between arms a fixed number of times.
(5) Section 6 shows how to generalize our results when one considers losses rather than gains, or signed games.
(6) Section 7 considers a framework fundamentally different from the previous sections, namely the stochastic multi-armed bandit problem. There we propose a new forecaster, MOSS, for which we prove an optimal distribution-free rate as well as a logarithmic distribution-dependent rate.
(7) Finally Section 9 contains most of the proofs.

2. The implicitly normalized forecaster

In this section, we define a new class of randomized policies for the general prediction game. Let us consider a continuously differentiable function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ satisfying

$$\psi' > 0, \qquad \lim_{x \to -\infty} \psi(x) < 1/K, \qquad \lim_{x \to 0} \psi(x) \geqslant 1. \tag{3.1}$$

LEMMA 3.1. *There exists a continuously differentiable function $C : \mathbb{R}_+^K \to \mathbb{R}$ satisfying for any $x = (x_1, \ldots, x_K) \in \mathbb{R}_+^K$,*

$$\max_{i=1,\ldots,K} x_i < C(x) \leqslant \max_{i=1,\ldots,K} x_i - \psi^{-1}\left(1/K\right), \tag{3.2}$$

and

$$\sum_{i=1}^{K} \psi(x_i - C(x)) = 1. \tag{3.3}$$

PROOF. Consider a fixed $x = (x_1, \ldots, x_K)$. The decreasing function $\varphi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ satisfies

$$\lim_{c \to \max_{i=1,\ldots,K} x_i} \varphi(c) > 1 \qquad \text{and} \qquad \lim_{c \to +\infty} \varphi(c) < 1.$$

From the intermediate value theorem, there is a unique $C(x)$ satisfying $\varphi(C(x)) = 1$. From the implicit function theorem, the mapping $x \mapsto C(x)$ is continuously differentiable. □

The implicitly normalized forecaster (INF) is defined in Figure 2. Equality (3.3) makes the fourth step in Figure 2 legitimate. From (3.2), $C(V_t)$ is roughly equal to $\max_{i=1,\ldots,K} V_{i,t}$. Recall that $V_{i,t}$ is an estimate of the cumulative gains for arm i. This means that INF chooses the probability assigned to arm i as a function of the (estimated) regret. Note that, in spirit, it is similar to the traditional weighted average forecaster, see e.g. Section 2.1 of Cesa-Bianchi and Lugosi [2006], where the probabilities are proportional to a function of the difference between the (estimated) cumulative reward of arm i and the cumulative reward of the policy, which should be, for a well-performing policy, of order $C(V_t)$. However, the normalization by division (that weighted

INF (Implicitly Normalized Forecaster):

Parameters:

- the continuously differentiable function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ satisfying (3.1)
- the estimates $v_{i,t}$ of $g_{i,t}$ based on the (drawn arms and) observed rewards at time t (and before time t)

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

For each round $t = 1, 2, \ldots,$

(1) Draw an arm I_t from the probability distribution p_t.

(2) Use the observed reward(s) to build the estimate $v_t = (v_{1,t}, \ldots, v_{K,t})$ of $(g_{1,t}, \ldots, g_{K,t})$ and let: $V_t = \sum_{s=1}^t v_s = (V_{1,t}, \ldots, V_{K,t})$.

(3) Compute the normalization constant $C_t = C(V_t)$.

(4) Compute the new probability distribution $p_{t+1} = (p_{1,t+1}, \ldots, p_{K,t+1})$ where

$$p_{i,t+1} = \psi(V_{i,t} - C_t).$$

Figure 2: The proposed policy for the general prediction game.

average forecasters perform) is fundamentally different from the normalization by shift of the real axis (that INF performs). Nonetheless, we can recover exactly the exponentially weighted average forecaster of Cesa-Bianchi and Lugosi [2006] because of the special relation of the exponential with the addition and the multiplication.

- Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0, 1)$. Then conditions (3.1) are clearly satisfied and equation (3.3) is equivalent to

$$\exp(-\eta C(x)) = \frac{1 - \gamma}{\sum_{i=1}^K \exp(\eta x_i)},$$

which implies

$$p_{i,t+1} = (1 - \gamma) \frac{\exp(\eta V_{i,t})}{\sum_{j=1}^K \exp(\eta V_{i,t})} + \frac{\gamma}{K}.$$

In other words, for the full information case (label efficient or not), we recover the exponentially weighted average forecaster (with $\gamma = 0$) while for the bandit game we recover EXP3. For the bandit label efficient game, it does not give us the GREEN policy proposed in Allenberg et al. [2006] but rather the straightforward modification of the exponentially weighted average forecaster to this game. Theorem 3.5 below gives a unified view on this algorithm for these four games. In particular, we recover all the upper bounds (up to constant factors) of Theorems 3.1, 3.2, 3.3 and 3.4. In the following, we will refer to this algorithm as the "exponentially weighted average forecaster" whatever the game is.

- Another fruitful choice of the function ψ is $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q > 1, \eta > 0$ and $\gamma \in [0, 1)$. Obviously, it also satisfies conditions (3.1). We will refer to this strategy as the "polynomial INF". Here the (normalizing) function C has no closed form expression (this is a consequence of Abel's impossibility theorem). Actually this remark holds in general, hence the name of the general policy. However this does not lead to a major computational issue since, in the interval given by (3.2), $C(x)$ is the unique solution of $\varphi(c) = 1$, where $\varphi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ is a decreasing function. We will prove that

the polynomial INF forecaster generates nicer probability updates than the exponentially weighted average forecasteras as, for bandits games (label efficient or not), it allows to remove the extraneous $\log K$ factor in the pseudo-regret bound.

Our main contribution is a uniform treatment of the Implicitly Normalized Forecaster for different functions ψ and different estimates of the gains $g_{i,t}$, hence for different prediction games. The general statement, Theorem 3.20, can be found in Section 8. The proof starts with an Abel transformation and consequently is "orthogonal" to the usual argument. Then we use a Taylor-Lagrange expansion and technical arguments to control the residual terms, see Section 8 for the details. We propose here the specialization of Theorem 3.20 to the two functions we discussed above.

THEOREM 3.5 (General regret bound for Exp INF). *Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0,1)$. Let $(v_{i,t})_{1 \leqslant i \leqslant K,\, 1 \leqslant t \leqslant n}$ be a sequence of nonnegative real numbers,*

$$B_t = \max_{1 \leqslant i < j \leqslant K} |v_{i,t} - v_{j,t}|, \quad \text{and} \quad B = \max_t B_t.$$

If $\gamma = 0$ then INF satisfies:

$$\max_{1 \leqslant i \leqslant K} \sum_{t=1}^{n} v_{i,t} - \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t} \leqslant \frac{\log K}{\eta} + \frac{\eta}{2} \exp(2\eta B) \sum_{t=1}^{n} B_t^2. \tag{3.4}$$

Moreover if

$$\frac{\gamma}{K} \geqslant \frac{\eta \exp(2B\eta)\left[1 + \exp(2B\eta)\right]}{2} \max_{i,t} p_{i,t} v_{i,t}, \tag{3.5}$$

then INF satisfies:

$$\max_{1 \leqslant i \leqslant K} \sum_{t=1}^{n} v_{i,t} - \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t} \leqslant \frac{1}{\eta} \log\left(\frac{K}{1-\gamma}\right) + \frac{\gamma}{1-\gamma} \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t}. \tag{3.6}$$

In the above Theorem, it is always trivial to upper bound the right hand side of (3.4) and (3.6). Thus, to derive concrete bounds from it, most of the work lies in relating the left hand side with the different notions of regret we consider. Note that in the case of the pseudo-regret this task is also trivial. On the other hand for high probability regret bounds, we will use concentration inequalities on top of (3.4) and (3.6). Expected regret bounds are then obtained by integration of the high probability bounds.

THEOREM 3.6 (General regret bound for Poly INF). *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q > 1, \eta > 0$ and $\gamma \in [0,1)$. Let $(v_{i,t})_{1 \leqslant i \leqslant K,\, 1 \leqslant t \leqslant n}$ be a sequence of nonnegative real numbers,*

$$B_t = \max_{1 \leqslant i < j \leqslant K} |v_{i,t} - v_{j,t}|, \quad \text{and} \quad B = \max_t B_t.$$

If $\gamma = 0$ then INF satisfies:

$$\max_{1 \leqslant i \leqslant K} \sum_{t=1}^{n} v_{i,t} - \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t} \leqslant \frac{q}{q-1} \eta K^{1/q} + \frac{q}{2\eta} \exp\left(2\frac{q+1}{\eta} B\right) \sum_{t=1}^{n} B_t^2. \tag{3.7}$$

Moreover if $v_{i,t} = \frac{c_t}{p_{i,t}} \mathbb{1}_{i=I_t}$ where c_t is a random variable independent of everything else, and such that $\mathbb{E} c_t \leqslant 1$, $\exists c > 0 : c_t \in [0,c]$ and $q\eta/c > (\frac{(q-1)K}{\gamma})^{(q-1)/q}$, then

$$\mathbb{E}\left(\max_{1 \leqslant i \leqslant K} \sum_{t=1}^{n} v_{i,t} - \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t}\right)$$

$$\leqslant \frac{1}{1-\gamma}\left\{\gamma n + \frac{q}{q-1}\eta K^{\frac{1}{q}} + \frac{\gamma n}{(q-1)K}\left(\frac{(q-1)cK\mu^2(1+\mu^2)}{2\gamma\eta}\right)^q\right\}, \tag{3.8}$$

where

$$\mu = \exp\left\{\frac{(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1 - \frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

Note that (3.8) is weaker than its counterpart (3.6) in the sense that it holds only in expectation. In fact, as we will see in the proof of Theorem 3.15, one can also prove a bound of the same form without the expectation. However, the general statement is less readable, hence we decided to provide here only the bound in expectation.

3. The full information (FI) game

The purpose of this section is to show how to use the Implicitly Normalized Forecaster in order to recover known minimax bounds (up to constant factors) in the full information game. In this section, we set $v_{i,t} = g_{i,t}$, which is possible since the rewards for all arms are observed in the full information setting. The proofs of Theorem 3.7 and Theorem 3.8 trivially follows from Theorem 3.5 and Theorem 3.6. We start with $\psi(x) = \exp(\eta x)$ for which INF reduces to the exponentially weighted average forecaster.

THEOREM 3.7 (Exponentially weighted average forecaster in the FI game). *Let $\psi(x) = \exp(\eta x)$ with $\eta > 0$. Let $v_{i,t} = g_{i,t}$. Then in the full information game, INF satisfies:*

$$\max_{1\leqslant i\leqslant K}\sum_{i=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \leqslant \frac{\log K}{\eta} + \exp(2\eta)\frac{\eta n}{2}. \tag{3.9}$$

In particular with $\eta = \sqrt{\frac{\log K}{n}}$ we get

$$\mathbb{E}R_n \leqslant \sqrt{5n\log K}$$

REMARK 3.1. *This result has to be compared with Corollary 2.2 of Cesa-Bianchi and Lugosi [2006] which derives the bound $\frac{\log K}{\eta} + \frac{\eta n}{2}$ from a more general theorem which can apply to other forecasters. Moreover it is proved in Theorem 2.2 of Cesa-Bianchi and Lugosi [2006] that the optimal bound is $\frac{\log K}{\eta} + \frac{\eta n}{8}$ (there the proof is specifically tailored to the exponential). While our bound has the same shape (η has to be thought as small, that is $\exp(2\eta)$ is close to 1), it does not capture the best constant. It seems to be the price for having a general Theorem applying to a large class of forecasters and various games.*

Now we consider a new algorithm for the FI game, that is INF with $\psi(x) = \left(\frac{\eta}{-x}\right)^q$.

THEOREM 3.8 (Polynomial INF in the FI game). *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q$ with $\eta > 0$ and $q > 1$. Let $v_{i,t} = g_{i,t}$. Then in the full information game, INF satisfies:*

$$\max_{1\leqslant i\leqslant K}\sum_{i=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \leqslant \frac{q}{q-1}\,\eta K^{1/q} + \exp\left(\frac{4q}{\eta}\right)\frac{qn}{2\eta}. \tag{3.10}$$

In particular with $q = 3\log K$ and $\eta = 1.8\sqrt{n\log K}$ we get

$$\mathbb{E}R_n \leqslant 7\sqrt{n\log K}.$$

REMARK 3.2. *By using the Hoeffding-Azuma inequality, see Theorem 10.1, one can derive high probability bounds from* (3.9) *and* (3.10)*: for instance, from* (3.10)*, for any $\delta > 0$, with*

probability at least $1-\delta$*, the polynomial INF satisfies:*

$$R_n \leqslant \frac{q}{q-1}\eta K^{1/q} + \exp\left(\frac{4q}{\eta}\right)\frac{qn}{2\eta} + \sqrt{\frac{n\log(\delta^{-1})}{2}}.$$

4. The limited feedback games

4.1. Label efficient game (LE).

4.1.1. *Soft constraint on the number of queried reward vectors.* As in Section 3, the purpose of this section is to show how to use the Implicitly Normalized Forecaster in order to recover known minimax bounds (up to constant factors) in a slight modification of the LE game.

Let us consider the following policy. At each round, we draw a Bernoulli random variable Z_t, with parameter $\varepsilon = m/n$, to decide whether we ask for the gains or not. Note that we do not fulfill exactly the requirement of the game as we might ask a bit more than m reward vectors. We do so in order to avoid technical details and focus on the main argument of the proof. The exact problem will be addressed in Section 4.1.2, where, in addition, we will prove bounds on the expected regret $\mathbb{E}R_n$ instead of just the pseudo-regret $\overline{R}_n$.

In this section, the estimate of $g_{i,t}$ is $v_{i,t} = \frac{g_{i,t}}{\varepsilon}Z_t$, which is observable since the rewards at time t for all arms are observed when $Z_t = 1$.

THEOREM 3.9 (Exponentially weighted average forecaster in the LE game). *Let* $\psi(x) = \exp(\eta x)$ *with* $\eta = \frac{\sqrt{m\log K}}{n}$*. Let* $v_{i,t} = \frac{g_{i,t}}{\varepsilon}Z_t$ *with* $\varepsilon = \frac{m}{n}$*. Then in the (soft) LE game, INF satisfies:*

$$\overline{R}_n \leqslant n\sqrt{\frac{5\log K}{m}}.$$

A similar result can be proved for the INF forecaster with $\psi(x) = \left(\frac{\eta}{-x}\right)^q$, $\eta > 0$ and q of order $\log K$. We do not state it since we will prove a stronger result in the next section.

4.1.2. *Hard constraint on the number of queried reward vectors.* The goal of this section is to push the idea that by using appropriate high probability bounds, one can control the expected regret $\mathbb{E}R_n = \mathbb{E}\max_{i=1,\dots,K}\sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$ instead of just the pseudo-regret $\overline{R}_n = \max_{i=1,\dots,K}\mathbb{E}\sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$. Most previous works have obtained results for $\overline{R}_n$. These results are interesting for oblivious opponents, that is when the adversary's choices of the rewards do not depend on the past draws and obtained rewards, since in this case a uniform (over all oblivious adversaries) bound on the pseudo-regret $\overline{R}_n$ implies the same bound on the expected regret $\mathbb{E}R_n$. This follows from noting that the expected regret against an oblivious adversary is smaller than the maximal pseudo-regret over deterministic adversaries. For non-oblivious opponents, upper bounds on $\overline{R}_n$ are rather weak statements and high probability bounds on R_n or bounds on $\mathbb{E}R_n$ are desirable. In Auer [2002], Cesa-Bianchi and Lugosi [2006], high probability bounds on R_n have been given. Unfortunately, the policies proposed there are depending on the confidence level of the bound. As a consequence, the resulting best bound on $\mathbb{E}R_n$, obtained by choosing the policies with confidence level parameter of order $1/n$, has an extraneous $\log n$ term. Specifically, from Theorem 6.2 of Cesa-Bianchi and Lugosi [2006], one can immediately derive $\mathbb{E}R_n \leqslant 8n\sqrt{\frac{\log(4K)+\log(n)}{m}} + 1$. The theorems of this section essentially show that the $\log n$ term can be removed.

As in Section 4.1.1, we still use a draw of a Bernoulli random variable Z_t to decide whether we ask for the gains or not. The difference is that, if $\sum_{s=1}^{t-1} Z_s \geqslant m$, we do not ask for the gains (as we are not allowed to do so). To avoid that this last constraint interferes in the analysis, the parameter

of the Bernoulli random variable is set to $\varepsilon = \frac{3m}{4n}$ and the probability of the event $\sum_{t=1}^n Z_t > m$ is upper bounded. The estimate of $g_{i,t}$ remains $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$.

THEOREM 3.10 (Exponentially weighted average forecaster in the LE game). *Let $\psi(x) = \exp(\eta x)$ with $\eta = \frac{\sqrt{m \log K}}{n}$. Let $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$ with $\varepsilon = \frac{3m}{4n}$. Then in the LE game, for any $\delta > 0$, with probability at least $1 - \delta$, INF satisfies:*

$$R_n \leqslant 2n\sqrt{\frac{\log K}{m}} + n\sqrt{\frac{27 \log(2K\delta^{-1})}{m}},$$

and

$$\mathbb{E}R_n \leqslant 8n\sqrt{\frac{\log(6K)}{m}}.$$

THEOREM 3.11 (Polynomial INF in the LE game). *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q$ with $q = 3\log(6K)$ and $\eta = 2n\sqrt{\frac{\log(6K)}{m}}$. Let $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$ with $\varepsilon = \frac{3m}{4n}$. Then in the LE game, for any $\delta > 0$, with probability at least $1 - \delta$, INF satisfies:*

$$R_n \leqslant 5n\sqrt{\frac{\log K}{m}} + n\sqrt{\frac{27 \log(2K\delta^{-1})}{m}},$$

and

$$\mathbb{E}R_n \leqslant 11n\sqrt{\frac{\log(6K)}{m}}.$$

4.2. Bandit game. This section is cut into two parts. In the first one, from Theorem 3.5 and Theorem 3.6, we derive upper bounds on $\overline{R}_n = \max_{i=1,\dots,K} \mathbb{E}\sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$. To bound $\mathbb{E}R_n = \mathbb{E}\max_{i=1,\dots,K} \sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$, we will then use high probability bounds on top of the use of these theorems. Since this makes the proofs more intricate, we have chosen to provide the less general results, but easier to obtain, in Section 4.2.1 and the more general ones in Section 4.2.2.

The main results here are that, by using the INF with a polynomial function ψ, we obtain an upper bound of order $\sqrt{nK}$ for $\overline{R}_n$, which imply a bound of order $\sqrt{nK}$ on $\mathbb{E}R_n$ for oblivious adversaries (see the reasoning at the beginning of Section 4.1.2). In the general case (containing the non-oblivious opponent), we show an upper bound of order $\sqrt{nK \log K}$ on $\mathbb{E}R_n$. We conjecture that this bound cannot be improved, that is the opponent may take advantage of the past to make the player pay a regret with the extra logarithmic factor (see Remark 3.3).

4.2.1. *Bounds on the pseudo-regret.* In this section, the estimate of $g_{i,t}$ is $v_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$, which is observable since the reward $g_{I_t,t}$ is revealed at time t.

THEOREM 3.12 (Exponentially weighted average forecaster in the bandit game). *Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{K\log(2K)}{n}}\right)$ and $\eta = \sqrt{\frac{\log(2K)}{9nK}}$. Let $v_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$. Then in the bandit game, INF satisfies:*

$$\overline{R}_n \leqslant \sqrt{20\, nK \log(2K)}.$$

THEOREM 3.13 (Polynomial INF in the bandit game). *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\gamma = \sqrt{K/n}$, $\eta = 3\sqrt{n}$ and $q = 2$. Let $v_{i,t} = \frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{I_t=i}$. Then in the bandit game, INF satisfies:*

$$\overline{R}_n \leqslant 10\sqrt{nK}.$$

We have arbitrarily chosen $q = 2$ to provide an explicit upper bound. More generally, it is easy to check that for any real number $q > 1$, we obtain the convergence rate $\sqrt{nK}$, provided that γ and η are respectively taken of order $\sqrt{K/n}$ and $\sqrt{nK}/K^{1/q}$.

4.2.2. *High probability bounds and bounds on the expected regret.* Theorems 3.12 and 3.13 provide upper bounds on $\overline{R}_n = \max_{i=1,\dots,K} \mathbb{E}\sum_{t=1}^n \big(g_{i,t}-g_{I_t,t}\big)$. To bound $\mathbb{E}R_n = \mathbb{E}\max_{i=1,\dots,K}\sum_{t=1}^n \big(g_{i,t}-g_{I_t,t}\big)$, we will use high probability bounds. First we need to modify the estimates of $g_{i,t}$ by considering $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i} + \frac{\beta}{p_{i,t}}$ with $0 < \beta \leqslant 1$, as was proposed in Auer [2002][3].

THEOREM 3.14 (Exponentially weighted average forecaster in the bandit game). *Let* $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ *with* $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{2K\log(2K)}{n}}\right)$ *and* $\eta = \sqrt{\frac{\log(2K)}{8nK}}$. *Let* $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i} + \frac{\beta}{p_{i,t}}$ *with* $\beta = \sqrt{\frac{\log(2K)}{nK}}$. *Then in the bandit game, for any* $\delta > 0$, *with probability at least* $1-\delta$, *INF satisfies:*

$$R_n \leqslant 7\sqrt{nK\log(2K)} + \sqrt{\frac{nK}{\log(2K)}}\log(\delta^{-1}),$$

and

$$\mathbb{E}R_n \leqslant 8\sqrt{nK\log(2K)}.$$

This theorem is similar to Theorem 6.10 of Cesa-Bianchi and Lugosi [2006]. The main difference here is that the high probability bound holds for any confidence level, and not only for a confidence level depending on the algorithm. As a consequence, our algorithm, unlike the one proposed in previous works, satisfies both a high probability bound and an expected regret bound of order $\sqrt{nK\log(K)}$.

THEOREM 3.15 (Polynomial INF in the bandit game). *Let* $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ *with* $\gamma = \sqrt{\frac{K}{n}}$, $\eta = 3\sqrt{n}$ *and* $q = 2$. *Let* $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i} + \frac{\beta}{p_{i,t}}$ *with* $\beta = \sqrt{\frac{1}{nK}}$. *Then in the bandit game, against an oblivious adversary, for any* $\delta > 0$, *with probability at least* $1-\delta$, *INF satisfies:*

$$R_n \leqslant 10.7\sqrt{nK} + 1.5\sqrt{nK}\log(\delta^{-1}), \tag{3.11}$$

and

$$\mathbb{E}R_n \leqslant 12.2\sqrt{nK}.$$

Moreover in the general case (containing the non-oblivious opponent), with $\beta = \sqrt{\frac{\log(2K)}{nK}}$, *it satisfies with probability at least* $1-\delta$,

$$R_n \leqslant 8.6\sqrt{nK} + 2.1\sqrt{nK\log(2K)} + 1.5\sqrt{\frac{nK}{\log(2K)}}\log(\delta^{-1}) \tag{3.12}$$

and

$$\mathbb{E}R_n \leqslant 10\sqrt{nK} + 2.1\sqrt{nK\log(2K)}.$$

REMARK 3.3. *We conjecture that the bound* $\sqrt{nK\log K}$ *on* $\mathbb{E}R_n$ *cannot be improved in the general case containing the non-oblivious opponent. Here is the main argument to support our conjecture. Consider an adversary choosing all rewards to be equal to one until time* $n/2$ *(say* n *is even to simplify). Then, let* $\hat{k}$ *denote the arm for which the estimate* $V_{i,n/2} = \sum_{1\leqslant t\leqslant n/2}\frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$ *of the cumulative reward of arm* i *is the smallest. After time* $n/2$, *all rewards are chosen to be equal to zero except for arm* $\hat{k}$ *for which the rewards are still chosen to be equal to* 1. *Since it can be proved that* $\max_{i\in\{1,\dots,K\}} V_{i,n/2} - \min_{j\in\{1,\dots,K\}} V_{j,n/2} \geqslant c\sqrt{nK\log K}$ *for some small enough* $c > 0$, *it seems that the INF algorithm achieving a bound of order* $\sqrt{nK}$ *on* $\mathbb{E}R_n$ *in the oblivious*

[3]The technical reason for this modification, which may appear surprising as it introduces a bias in the estimate of $g_{i,t}$, is that it allows to have high probability upper bounds with the correct rate on the difference $\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n v_{i,t}$. A second reason for this modification (but useless for this particular section) is that it allows to track the best expert (see Section 5).

setting will suffer an expected regret of order at least $\sqrt{nK\log K}$. While this does not prove the conjecture as one can design other algorithms, it makes the conjecture likely to hold.

4.3. Label efficient and bandit game (LE bandit). We consider the following policy. At each round, we draw a Bernoulli random variable Z_t, with parameter $\varepsilon = m/n$, to decide whether the gain of the chosen arm is revealed or not. Note that we do not fulfil exactly the requirement of the game as we might ask a bit more than m rewards (but, as was argued in Section 4.1.2, this is just a technical detail that requires some extra computations to be taken into account). In this section, the estimate of $g_{i,t}$ is $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}\frac{Z_t}{\varepsilon}$.

THEOREM 3.16 (Exponentially weighted average forecaster in the LE bandit game). *Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{K\log(2K)}{m}}\right)$ and $\eta = \frac{1}{n}\sqrt{\frac{m\log(2K)}{9K}}$. Let $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}\frac{Z_t}{\varepsilon}$ with $\varepsilon = \frac{m}{n}$. Then in the LE bandit game, INF satisfies:*

$$\overline{R}_n \leqslant n\sqrt{\frac{20K\log(2K)}{m}}.$$

THEOREM 3.17 (Polynomial INF in the LE bandit game). *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\gamma = \sqrt{\frac{K}{m}}$, $\eta = \frac{3n}{K^{1/q}}\sqrt{\frac{K}{m}}$ and $q = 2$. Let $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}\frac{Z_t}{\varepsilon}$ with $\varepsilon = \frac{m}{n}$. Then in the LE bandit game, INF satisfies:*

$$\overline{R}_n \leqslant 10n\sqrt{\frac{K}{m}}.$$

5. Tracking the best expert in the bandit game

In the previous sections, the cumulative gain of the forecaster was compared to the cumulative gain of the best single expert. Here, it will be compared to more flexible strategies that are allowed to switch actions. We will use the same algorithms as in Section 4.2.2, but with different parameters. We thus use

$$v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} + \frac{\beta}{p_{i,t}},$$

with $0 < \beta \leqslant 1$. The β term introduces a bias in the estimate of $g_{i,t}$, that constrains the differences $\max_{i=1,\dots,K} V_{i,t} - \min_{j=1,\dots,K} V_{j,t}$ to be relatively small. This is the key property in order to track the best switching strategy, provided that the number of switches is not too large. A switching strategy is defined by a vector $(i_1,\dots,i_n) \in \{1,\dots,K\}^n$. Its size is defined by

$$\mathcal{S}(i_1,\dots,i_n) = \sum_{t=1}^{n-1} \mathbb{1}_{i_{t+1}\neq i_t},$$

and its cumulative gain is

$$G_{(i_1,\dots,i_n)} = \sum_{t=1}^{n} g_{i_t,t}.$$

The regret of a forecaster with respect to the best switching strategy with S switches is then given by:

$$R_n^S = \max_{(i_1,\dots,i_n):\,\mathcal{S}(i_1,\dots,i_n)\leqslant S} G_{(i_1,\dots,i_n)} - \sum_{t=1}^{n} g_{I_t,t}.$$

THEOREM 3.18 (INF for tracking the best expert in the bandit game). *Let $s = S\log\left(\frac{enK}{S}\right) + \log(2K)$ with $e = \exp(1)$ and the natural convention $S\log(enK/S) = 0$ for $S = 0$. Let $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} + \frac{\beta}{p_{i,t}}$ with $\beta = 2\sqrt{\frac{s}{nK}}$. Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{Ks}{n}}\right)$ and*

$\eta = \sqrt{\frac{s}{20nK}}$. *Then in the bandit game, for any* $0 \leqslant S \leqslant n-1$, *for any* $\delta > 0$, *with probability at least* $1-\delta$, *INF satisfies:*

$$R_n^S \leqslant 9\sqrt{nKs} + \sqrt{\frac{nK}{s}}\log(\delta^{-1}),$$

and

$$\mathbb{E}R_n^S \leqslant 10\sqrt{nKs}.$$

Note that for $S = 0$, we have $R_n^S = R_n$, and we recover (up to a constant) the result of Theorem 3.14.

REMARK 3.4. *Up to constant factors, the same bounds as the ones of Theorem 3.18 can be obtained (via a tedious proof not requiring new arguments than the ones presented in this work) for the INF forecaster using* $\psi(x) = \frac{c_1}{K}\left(\frac{\sqrt{snK}}{-x}\right)^{c_3 s} + c_2\sqrt{\frac{s}{nK}}$, *with* $s = S\log\left(\frac{enK}{S}\right) + \log(2K)$ *and appropriate constants* c_1, c_2 *and* c_3.

6. Gains vs losses, unsigned games vs signed games

To simplify, we have considered so far that the rewards were in $[0,1]$. Here is a trivial argument which shows how to transfer our analysis to loss games (i.e., games with only non-positive rewards), and more generally to signed games (i.e., games in which the rewards can be positive and negative). If the rewards, denoted now $g'_{i,t}$, are in some interval $[a,b]$ potentially containing zero, we set $g_{i,t} = \frac{g'_{i,t}-a}{b-a} \in [0,1]$. Then we can apply our analysis to:

$$\max_{i\in\{1,\dots,K\}}\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{I_t,t} = \frac{1}{b-a}\left(\max_{i\in\{1,\dots,K\}}\sum_{t=1}^n g'_{i,t} - \sum_{t=1}^n g'_{I_t,t}\right).$$

Note that a less straightforward analysis can be done by looking at the INF algorithm directly applied to the observed rewards (and not to the renormalized rewards). In this case, as it was already noted in Remark 6.5 of Cesa-Bianchi and Lugosi [2006], the behavior of the algorithm may be very different for loss and gain games. However it can be proved that our analysis still holds up to constants factors (one has to go over the proofs and make appropriate modifications).

7. Stochastic bandit game

By considering the deterministic case when the rewards are $g_{i,t} = 1$ if $i = 1$ and $g_{i,t} = 0$ otherwise, it can be proved that the INF policies considered in Theorem 3.12 and Theorem 3.13 have a pseudo-regret lower bounded by $\sqrt{nK}$. In this simple setting, and more generally in most of the stochastic multi-armed bandit problems, one would like to suffer a much smaller regret.

We recall that in the stochastic bandit considered in this section, the adversary samples the reward $g_{i,t}$ i.i.d from a fixed distribution ν_i on $[0,1]$ for each arm i. The suboptimality of an arm i is then measured by $\Delta_i = \max_{j=1,\dots,K}\mu_j - \mu_i$ where μ_i is the mean of ν_i. We provide now a strategy achieving a $\sqrt{nK}$ regret in the worst case, and a much smaller regret as soon as the Δ_i of the suboptimal arms are much larger than $\sqrt{K/n}$.

Let $\widehat{\mu}_{i,s}$ be the empirical mean of arm i after s draws of this arm. Let $T_i(t)$ denote the number of times we have drawn arm i on the first t rounds. In this section, we propose a policy, called MOSS (Minimax Optimal Strategy in the Stochastic case), inspired by the UCB1 policy (Auer et al., 2002). As in UCB1, each arm has an index measuring its performance, and at each round, we choose the arm having the highest index. The only difference with UCB1 is to use $\log\left(\frac{n}{Ks}\right)$ instead of $\log(t)$ at time t (see Figure 3). As a consequence, an arm that has been drawn more than n/K times has an index equal to the empirical mean of the rewards obtained from the arm, and

when it has been drawn close to n/K times, the logarithmic term is much smaller than the one of UCB1, implying less exploration of this already intensively drawn arm.

MOSS (Minimax Optimal Strategy in the Stochastic case):
For an arm i, define its index $B_{i,s}$ by

$$B_{i,s} = \widehat{\mu}_{i,s} + \sqrt{\frac{\max\left(\log(\frac{n}{Ks}), 0\right)}{s}}\,.$$

for $s \geqslant 1$ and $B_{i,0} = +\infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1)}$.

Figure 3: The proposed policy for the stochastic bandit game.

THEOREM 3.19. *Introduce* $\Delta = \min_{i\in\{1,\dots,K\}:\Delta_i>0} \Delta_i$. *MOSS satisfies*

$$\overline{R}_n \leqslant \frac{23K}{\Delta}\log\left(\max\left(\frac{110n\Delta^2}{K}, 10^4\right)\right), \tag{3.13}$$

and

$$\mathbb{E}R_n \leqslant 25\sqrt{nK}. \tag{3.14}$$

Besides, if there exists a unique arm with $\Delta_i = 0$, *we also have*

$$\mathbb{E}R_n \leqslant \frac{23K}{\Delta}\log\left(\max\left(\frac{140n\Delta^2}{K}, 10^4\right)\right). \tag{3.15}$$

The distribution-dependent bounds Inequalities (3.13) and (3.15) show the desired logarithmic dependence in n, while the distribution-free regret bound (3.14) has the minimax rate $\sqrt{nK}$.

REMARK 3.5. *The uniqueness of the optimal arm is really needed to have the logarithmic (in* n*) bound on the expected regret. This can be easily seen by considering a two-armed bandit in which both reward distributions are identical (and non degenerated). In this case, the pseudo-regret is equal to zero while the expected regret is of order* $\sqrt{n}$. *This reveals a fundamental difference between the expected regret and the pseudo-regret.*

REMARK 3.6. *A careful tuning of the constants in front and inside the logarithmic term of* $B_{i,s}$ *and of the thresholds used in the proof leads to smaller numerical constants in the previous theorem, and in particular to* $\sup \mathbb{E}R_n \leqslant 6\sqrt{nK}$. *However, it makes the proof more intricate. So we will only prove* (3.13).

8. General regret bound

THEOREM 3.20 (General regret bound for INF). *For any nonnegative real numbers* $v_{i,t}$, *where* $i \in \{1,\dots,K\}$ *and* $t \in \mathbb{N}^*$, *we still use* $v_t = (v_{1,t},\dots,v_{K,t})$ *and* $V_t = \sum_{s=1}^t v_t$. *Define* $[V_{t-1}, V_t] = \{\lambda V_{t-1} + (1-\lambda)V_t : \lambda \in [0,1]\}$. *Let*

$$B_t = \max_{1\leqslant i<j\leqslant K} |v_{i,t} - v_{j,t}|,$$

$$\rho = \max_{1\leqslant t\leqslant n} \max_{v,w\in[V_{t-1},V_t],\, 1\leqslant i\leqslant K} \frac{\psi'(v_i - C(v))}{\psi'(w_i - C(w))},$$

and

$$A_t = \min\left(B_t^2 \sum_{i=1}^K \psi' \circ \psi^{-1}(p_{i,t}), (1+\rho^2)\sum_{i=1}^K \psi' \circ \psi^{-1}(p_{i,t}) v_{i,t}^2\right).$$

Then the INF forecaster based on ψ satisfies:

(3.16)

$$\left(\max_{1\leqslant i\leqslant K} V_{i,n}\right) - \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \leqslant -\sum_{i=1}^K \left(p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du\right) + \frac{\rho^2}{2}\sum_{t=1}^n A_t.$$

PROOF. In the following we set $V_0 = 0 \in \mathbb{R}_+^K$ and $C_0 = C(V_0)$. The proof is divided into four steps.

First step: Rewriting $\sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t}$.

We start with a simple Abel transformation:

$$\begin{aligned}
\sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} &= \sum_{t=1}^n \sum_{i=1}^K p_{i,t}(V_{i,t} - V_{i,t-1}) \\
&= \sum_{i=1}^K p_{i,n+1} V_{i,n} + \sum_{i=1}^K \sum_{t=1}^n V_{i,t}(p_{i,t} - p_{i,t+1}) \\
&= \sum_{i=1}^K p_{i,n+1}\left(\psi^{-1}(p_{i,n+1}) + C_n\right) + \sum_{i=1}^K \sum_{t=1}^n (\psi^{-1}(p_{i,t+1}) + C_t)(p_{i,t} - p_{i,t+1}) \\
&= C_n + \sum_{i=1}^K p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \sum_{i=1}^K \sum_{t=1}^n \psi^{-1}(p_{i,t+1})(p_{i,t} - p_{i,t+1})
\end{aligned}$$

where the last step comes from the fact that $\sum_{i=1}^K p_{i,t} = 1$.

Second step: A Taylor-Lagrange expansion.

For $x \in [0,1]$ we define $f(x) = \int_0^x \psi^{-1}(u)du$. Remark that $f'(x) = \psi^{-1}(x)$ and $f''(x) = 1/\psi'(\psi^{-1}(x))$. Then by the Taylor-Lagrange formula, we know that for any i, there exists $\widetilde{p}_{i,t+1} \in [p_{i,t}, p_{i,t+1}]$ (with the convention $[a,b] = [b,a]$ when $a > b$) such that

$$f(p_{i,t}) = f(p_{i,t+1}) + (p_{i,t} - p_{i,t+1})f'(p_{i,t+1}) + \frac{(p_{i,t} - p_{i,t+1})^2}{2} f''(\widetilde{p}_{i,t+1}),$$

or, in other words:

$$(p_{i,t} - p_{i,t+1})\psi^{-1}(p_{i,t+1}) = \int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du - \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'(\psi^{-1}(\widetilde{p}_{i,t+1}))}.$$

Now by summing over t the first term on the right hand side becomes $\int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$. Moreover, since $x \to \psi(x - C(x))$ is continuous, there exists $W^{(i,t)} \in [V_t, V_{t+1}] \subset \mathbb{R}^K$ such that $\psi\left(W_i^{(i,t)} - C(W^{(i,t)})\right) = \widetilde{p}_{i,t+1}$. Thus we have

$$\sum_{i=1}^K \sum_{t=1}^n \psi^{-1}(p_{i,t+1})(p_{i,t} - p_{i,t+1}) = \sum_{i=1}^K \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du - \sum_{i=1}^K \sum_{t=1}^n \frac{(p_{i,t} - p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)} - C(W^{(i,t)})\right)}.$$

From the equality obtained in the first step, it gives

$$C_n - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} \leqslant -\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right) + \sum_{i=1}^{K}\sum_{t=1}^{n}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)} - C(W^{(i,t)})\right)}.$$

Third step: The mean value theorem to compute $(p_{i,t+1}-p_{i,t})^2$.

It is now convenient to consider the functions f_i and h_i defined for any $x \in \mathbb{R}_+^K$ by

$$f_i(x) = \psi(x_i - C(x)) \quad \text{and} \quad h_i(x) = \psi'(x_i - C(x)).$$

We are going to bound $p_{i,t+1} - p_{i,t} = f_i(V_t) - f_i(V_{t-1})$ by using the mean value theorem. To do so we need to compute the gradient of f_i. First, we have

$$\frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{1}_{i=j} - \frac{\partial C}{\partial x_j}(x)\right)h_i(x).$$

Now, by definition of C, we have $\sum_{k=1}^K f_k(x) = 1$ and thus $\sum_{k=1}^K \frac{\partial f_k}{\partial x_j}(x) = 0$, which implies

$$\frac{\partial C}{\partial x_j}(x) = \frac{h_j(x)}{\sum_{k=1}^K h_k(x)} \qquad \text{and} \qquad \frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{1}_{i=j} - \frac{h_j(x)}{\sum_{k=1}^K h_k(x)}\right)h_i(x).$$

Now the mean value theorem says that there exists $V^{(i,t)} \in [V_{t-1}, V_t]$ such that

$$f_i(V_t) - f_i(V_{t-1}) = \sum_{j=1}^{K} v_{j,t}\frac{\partial f_i}{\partial x_j}(V^{(i,t)}).$$

Thus we have

$$\begin{aligned}(p_{i,t}-p_{i,t+1})^2 &= \left(\sum_{j=1}^{K} v_{j,t}\left(\mathbb{1}_{i=j} - \frac{h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)h_i(V^{(i,t)})\right)^2 \\ &= h_i(V^{(i,t)})^2\left(v_{i,t} - \frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2.\end{aligned}$$

Fourth step: An almost variance term.

We introduce $\rho = \max_{v,w\in[V_{t-1},V_t],\, 1\leqslant t\leqslant n,\, 1\leqslant i\leqslant K} \frac{h_i(v)}{h_i(w)}$. Thus we have

$$\begin{aligned}\sum_{i=1}^{K}\sum_{t=1}^{n}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)} - C(W^{(i,t)})\right)} &= \sum_{i=1}^{K}\sum_{t=1}^{n}\frac{h_i(V^{(i,t)})^2}{2h_i(W^{(i,t)})}\left(v_{i,t} - \frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 \\ &\leqslant \frac{\rho^2}{2}\sum_{t=1}^{n}\sum_{i=1}^{K} h_i(V_{t-1})\left(v_{i,t} - \frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2.\end{aligned}$$

Now we need to control the term $\left(v_{i,t} - \frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2$. Remark that since the function ψ is increasing we know that $h_i(x) \geqslant 0, \forall x$. Now since $\forall i,j\ |v_{i,t} - v_{j,t}| \leqslant B_t$, we can simply bound this last term by B_t^2. A different bound can be obtained by using $(a-b)^2 \leqslant a^2 + b^2$ when

a and b have the same sign:

$$\begin{aligned}\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 &\leqslant v_{i,t}^2+\left(\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2\\ &\leqslant v_{i,t}^2+\frac{\sum_{j=1}^K v_{j,t}^2h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\\ &\leqslant v_{i,t}^2+\rho^2\frac{\sum_{j=1}^K v_{j,t}^2h_j(V_{t-1})}{\sum_{k=1}^K h_k(V_{t-1})}\end{aligned}$$

where the first inequality comes from the fact that both terms are nonnegative and the second inequality comes from Jensen's inequality. As a consequence, we have

$$\begin{aligned}\sum_{i=1}^K h_i(V_{t-1})\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 &\leqslant \sum_{i=1}^K h_i(V_{t-1})v_{i,t}^2+\rho^2\sum_{j=1}^K h_j(V_{t-1})v_{j,t}^2\\ &\leqslant (1+\rho^2)\sum_{i=1}^K h_i(V_{t-1})v_{i,t}^2\end{aligned}$$

We have so far proved

(3.17)

$$C_n-\sum_{t=1}^n\sum_{i=1}^K p_{i,t}v_{i,t}\leqslant -\sum_{i=1}^K\left(p_{i,n+1}\psi^{-1}(p_{i,n+1})+\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right)+\frac{\rho^2}{2}\sum_{t=1}^n A_t$$

The announced result is then obtained by using Inequality (3.2). □

LEMMA 3.2. *Let ψ be a convex function satisfying* (3.1) *and assume that there exists $B>0$ such that $\forall i,j,t\ |v_{i,t}-v_{j,t}|\leqslant B$. Then:*

$$\rho=\max_{1\leqslant t\leqslant n}\ \max_{v,w\in[V_{t-1},V_t],\,1\leqslant i\leqslant K}\frac{\psi'(v_i-C(v))}{\psi'(w_i-C(w))}\leqslant\sup_{x\in(-\infty,\psi^{-1}(1)]}\exp\left(B\frac{\psi''}{\psi'}(x)\right).$$

PROOF. Let $h_i(x)=\psi'(x_i-C(x)), m_i(x)=\psi''(x_i-C(x))$. For $\alpha\in[0,1]$ we note

$$\varphi(\alpha)=\log\left\{h_i(V_{t-1}+\alpha(V_t-V_{t-1}))\right\}.$$

Remark that we should rather note this function $\varphi_{i,t}(\alpha)$ but for sake of simplicity we omit this dependency. With these notations we have $\rho=\max_{\alpha,\beta\in[0,1];\,1\leqslant t\leqslant n,\,1\leqslant i\leqslant K}\exp(\varphi(\alpha)-\varphi(\beta))$. By the mean value theorem for any $\alpha,\beta\in[0,1]$ there exists $\xi\in[0,1]$ such that $\varphi(\alpha)-\varphi(\beta)=(\alpha-\beta)\varphi'(\xi)$. Now with the calculus done in the third step of the proof of Theorem 3.20 and using the notations $h_i:=h_i(V_{t-1}+\xi(V_t-V_{t-1}))$, $m_i:=m_i(V_{t-1}+\xi(V_t-V_{t-1}))$ we obtain

$$\varphi'(\xi)=\sum_{j=1}^K(V_{j,t}-V_{j,t-1})\left(\mathbb{1}_{i=j}-\frac{h_j}{\sum_{k=1}^K h_k}\right)\frac{m_i}{h_i}=\sum_{j=1}^K\frac{(v_{i,t}-v_{j,t})h_j}{\sum_{k=1}^K h_k}\frac{m_i}{h_i}.$$

Thus we get

$$|\varphi'(\xi)|\leqslant\max_{1\leqslant i,j\leqslant K}|v_{i,t}-v_{j,t}|\sup_{v\in[V_{t-1},V_t]}\frac{\psi''}{\psi'}\left(v_i-C(v)\right).$$

Moreover, using that $x\to\psi(x-C(x))$ is continuous we know that there exists $\widetilde{p}_{i,t+1}\in[p_{i,t},p_{i,t+1}]$ such that $\widetilde{p}_{i,t+1}=\psi(v_i-C(v))$ and thus $v_i-C(v)=\psi^{-1}(\widetilde{p}_{i,t+1})$. This concludes the proof. □

LEMMA 3.3. *Let ψ be a function satisfying* (3.1) *and assume that there exists $c>0$ such that $0\leqslant v_{i,t}\leqslant\frac{c}{p_{i,t}}\mathbb{1}_{i=I_t}$. We also assume that ψ'/ψ is a nondecreasing function and that there exists*

$a > 1$ such that $\psi\left(x + \frac{c}{\psi(x)}\right) \leqslant a\psi(x)$. Then:

$$\rho \leqslant \sup_{x \in (-\infty, \psi^{-1}(1)]} \exp\left(ac \frac{\psi''}{\psi \times \psi'}(x)\right).$$

PROOF. We extract from the previous proof that $\rho \leqslant \max_{\xi \in [0,1];\, 1 \leqslant t \leqslant n,\, 1 \leqslant i \leqslant K} \exp(|\varphi'(\xi)|)$ where

$$\varphi'(\xi) = \sum_{j=1}^{K} \frac{(v_{i,t} - v_{j,t}) h_j}{\sum_{k=1}^{K} h_k} \frac{m_i}{h_i}.$$

Note that, since the functions ψ and ψ'/ψ are nondecreasing, the function ψ is convex, hence $\psi'' \geqslant 0$ and $m_i \geqslant 0$. Now using our assumption on $v_{i,t}$ and since $p_{i,t} = f_i(V_{t-1})$, if $i \neq I_t$ we have:

$$|\varphi'(\xi)| = \frac{c \frac{h_{I_t}}{f_{I_t}(V_{t-1})}}{\sum_{k=1}^{K} h_k} \frac{m_i}{h_i}.$$

Noticing that for any x, y in $\mathbb{R}_-^*$, $\frac{\frac{\psi'}{\psi}(x) \times \frac{\psi''}{\psi'}(y)}{\psi'(x) + \psi'(y)} \leqslant \frac{\psi''(y)}{\psi'(y)\psi(y)}$, we obtain

$$|\varphi'(\xi)| \leqslant c \frac{f_{I_t}(V_{t-1} + \xi(V_t - V_{t-1}))}{f_{I_t}(V_{t-1})} \frac{m_i}{h_i \times f_i}$$

where we note $f_i := f_i(V_{t-1} + \xi(V_t - V_{t-1}))$.

On the other hand if $i = I_t$ then

$$|\varphi'(\xi)| \leqslant \frac{c}{f_i(V_{t-1})} \frac{m_i}{h_i}.$$

To finish we only have to prove that $f_{I_t}(V_{t-1} + \xi(V_t - V_{t-1})) \leqslant a f_{I_t}(V_{t-1})$. Since ψ is increasing it is enough to prove that $f_{I_t}(V_t) \leqslant a f_{I_t}(V_{t-1})$ which is equivalent to

$$\psi(V_{I_t,t-1} + v_{I_t,t} - C_t) \leqslant a\psi(V_{I_t,t-1} - C_{t-1}).$$

Since $0 \leqslant v_{i,t} \leqslant \frac{c}{p_{i,t}} \mathbb{1}_{i=I_t}$ and C is an increasing function in each of its argument it is enough to prove

$$\psi\left(V_{I_t,t-1} - C_{t-1} + \frac{c}{\psi(V_{I_t,t-1} - C_{t-1})}\right) \leqslant a\psi(V_{I_t,t-1} - C_{t-1})$$

which is true by hypothesis on ψ. □

LEMMA 3.4. *Let ψ be a function satisfying* (3.1) *and assume that $v_{i,t} = \frac{g_{i,t}\mathbb{1}_{i=I_t} + \beta}{p_{i,t}}$ with $g_{i,t} \in [0,1]$ and $p_{i,t} \geqslant \gamma/K$. We also assume that ψ'/ψ is a nondecreasing function and that there exists $a > 1$ such that $\psi\left(x + \frac{1+\beta}{\psi(x)}\right) \leqslant a\psi(x)$. Then:*

$$\rho \leqslant \sup_{x \in (-\infty, \psi^{-1}(1)]} \exp\left(a \frac{\psi''}{\psi \times \psi'}(x) + \frac{\beta K}{\gamma} \frac{\psi''}{\psi'}\right).$$

PROOF. We extract from the previous proof that $\rho \leqslant \max_{\xi \in [0,1];\, 1 \leqslant t \leqslant n,\, 1 \leqslant i \leqslant K} \exp(|\varphi'(\xi)|)$ where

$$\begin{aligned}
\varphi'(\xi) &= \sum_{j=1}^{K} \frac{(v_{i,t} - v_{j,t}) h_j}{\sum_{k=1}^{K} h_k} \frac{m_i}{h_i} \\
&= \sum_{j=1}^{K} \frac{\left(\frac{g_{i,t}}{p_{i,t}} \mathbb{1}_{i=I_t} - \frac{g_{j,t}}{p_{j,t}} \mathbb{1}_{j=I_t}\right) h_j}{\sum_{k=1}^{K} h_k} \frac{m_i}{h_i} + \sum_{j=1}^{K} \frac{\left(\frac{\beta}{p_{i,t}} - \frac{\beta}{p_{j,t}}\right) h_j}{\sum_{k=1}^{K} h_k} \frac{m_i}{h_i}.
\end{aligned}$$

For the second term we can apply the same reasoning than for Lemma 3.2 to bound it by $\sup_{x\in(-\infty,\psi^{-1}(1)]} \frac{\beta K}{\gamma}\frac{\psi''}{\psi'}(x)$ while for the first term we can use the proof of Lemma 3.3 to bound it by $\sup_{x\in(-\infty,\psi^{-1}(1)]} a\frac{\psi''}{\psi\times\psi'}(x)$ which ends the proof. □

9. Proofs

Proof of Theorem 3.5. We make use of Theorem 3.20 and start with straightforward computations to bound the first sum in (3.16). We have $\psi^{-1}(x) = \frac{1}{\eta}\log(x-\gamma/K)$ which admits as a primitive $\int \psi^{-1}(u)du = \frac{1}{\eta}\left[(u-\gamma/K)\log(u-\gamma/K) - u\right]$. Thus one immediately gets

$$-\int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du - p_{i,n+1}\psi^{-1}(p_{i,n+1}) \\ = \frac{1}{\eta}\left(\frac{1}{K} - \frac{1-\gamma}{K}\log\left(\frac{1-\gamma}{K}\right) - p_{i,n+1} - \frac{\gamma}{K}\log\left(p_{i,n+1} - \frac{\gamma}{K}\right)\right).$$

Summing over i proves that

$$-\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right) = \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right) - \frac{\gamma}{K}\sum_{i=1}^{K}\psi^{-1}(p_{i,n+1}).$$

With the notations of Theorem 3.20, we need now to bound ρ and A_t. For the former we use Lemma 3.2 which directly shows:

$$\rho \leqslant \exp(\eta B).$$

For the latter we distinguish two cases. If $\gamma = 0$ we use

$$A_t \leqslant B_t^2 \sum_{i=1}^{K} \psi'\circ\psi^{-1}(p_{i,t}) = \eta B_t^2,$$

which concludes the proof of (3.4). On the other hand if $\gamma > 0$ we use

$$A_t \leqslant (1+\rho^2)\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t})v_{i,t}^2 \leqslant (1+\rho^2)\eta\sum_{i=1}^{K}p_{i,t}v_{i,t}^2.$$

Now recall that, as seen in (3.17), Theorem 3.20 holds with the maximum replaced by C_n. Thus, if (3.5) is satisfied we have the following bound

$$\begin{aligned} &C_n - \sum_{t=1}^{n}\sum_{i=1}^{K}p_{i,t}v_{i,t} \\ &\leqslant \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right) - \frac{\gamma}{K}\sum_{i=1}^{K}\left(\sum_{t=1}^{n}v_{i,t} - C_n\right) + \frac{\eta\exp(2B\eta)\left[1+\exp(2B\eta)\right]}{2}\sum_{t=1}^{n}\sum_{i=1}^{K}p_{i,t}v_{i,t}^2 \\ &\leqslant \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right) + \gamma C_n, \end{aligned}$$

which concludes the proof of (3.6) by using Inequality (3.2).

Proof of Theorem 3.6. We make use of Theorem 3.20 and start with straightforward computations to bound the first sum in (3.16). We have $\psi^{-1}(x) = -\eta(x-\gamma/K)^{-1/q}$ which admits as a primitive

$\int \psi^{-1}(u)du = \frac{-\eta}{1-1/q}(u-\gamma/K)^{1-1/q}$. Thus one immediately gets

$$\int_{p_{i,n+1}}^{1/K}(-\psi^{-1})(u)du \leqslant \frac{\eta}{1-1/q}\frac{1}{K^{1-1/q}} - \eta(p_{i,n+1}-\gamma/K)^{1-1/q}$$

and

$$p_{i,n+1}(-\psi^{-1})(p_{i,n+1}) = -\frac{\gamma}{K}\psi^{-1}(p_{i,n+1}) + \eta(p_{i,n+1}-\gamma/K)^{1-1/q}.$$

Summing over i proves that

$$-\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right) \leqslant \frac{q}{q-1}\eta K^{1/q} - \frac{\gamma}{K}\sum_{i=1}^{K}\psi^{-1}(p_{i,n+1}).$$

With the notations of Theorem 3.20, we need now to bound ρ and A_t. First we deal with the case $\gamma = 0$. Lemma 3.2 implies $\rho \leqslant \exp(B(q+1)/\eta)$ since we have $\frac{\psi''}{\psi'}(x) = \frac{q+1}{-x} = \frac{q+1}{\eta}\psi(x)^{1/q}$. The proof of (3.7) is concluded by $\psi' = \frac{q}{\eta}\psi^{(q+1)/q}$, and

$$A_t \leqslant B_t^2\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t}) = B_t^2\sum_{i=1}^{K}\frac{q}{\eta}p_{i,t}^{(q+1)/q} \leqslant \frac{q}{\eta}B_t^2.$$

Unfortunately the case $\gamma > 0$ is much more intricate. In this case we consider a specific form for the estimates $v_{i,t}$, see the assumptions in Theorem 3.6. We start by using lemma 3.3 to prove that $\rho \leqslant \mu$. First we have $\frac{\psi''}{\psi'} = \frac{q+1}{\eta}(\psi-\gamma/K)^{1/q} \leqslant \frac{q+1}{\eta}\psi^{1/q}$. Besides, for any $a \geqslant b \geqslant c$ we have $\frac{a}{b} \leqslant \frac{a-c}{b-c}$ and thus for any $x < 0$, we have

$$\frac{\psi\left(x+\frac{c}{\psi(x)}\right)}{\psi(x)} \leqslant \frac{\psi\left(x+\frac{c}{\psi(x)}\right)-\frac{\gamma}{K}}{\psi(x)-\frac{\gamma}{K}} = \left(1-\frac{c}{-x\psi(x)}\right)^{-q} \leqslant \left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}.$$

Thus lemma 3.3 gives us

$$\rho \leqslant \exp\left\{\frac{(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\} = \mu. \tag{3.18}$$

Next we use $\psi' = \frac{q}{\eta}(\psi-\gamma/K)^{(q+1)/q}$ and the form of $v_{i,t}$ to get

$$A_t \leqslant (1+\rho^2)\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t})v_{i,t}^2 \leqslant \frac{q(1+\mu^2)}{\eta}\sum_{i=1}^{K}p_{i,t}^{(q+1)/q}v_{i,t}^2 \leqslant \frac{q(1+\mu^2)}{\eta}\sum_{i=1}^{K}p_{i,t}^{1/q}v_{i,t}c_t.$$

We note $\mathbb{E}_t$ for the expectation with respect to c_t and the random draw of I_t from p_t. Using that $c_t \leqslant c$, $\mathbb{E}_t v_{i,t} \leqslant 1$ and Hölder's inequality, we obtain

$$\begin{aligned}\mathbb{E}A_t \leqslant \frac{qc(1+\mu^2)}{\eta}\mathbb{E}\sum_{i=1}^{K}p_{i,t}^{1/q}\mathbb{E}_t v_{i,t} &\leqslant \frac{qc(1+\mu^2)}{\eta}\mathbb{E}\left(\sum_{i=1}^{K}(\mathbb{E}_t v_{i,t})^{q/(q-1)}\right)^{(q-1)/q} \\ &\leqslant \frac{qc(1+\mu^2)}{\eta}\mathbb{E}\left(\sum_{i=1}^{K}\mathbb{E}_t v_{i,t}\right)^{(q-1)/q}.\end{aligned}$$

Now recall that, as seen in (3.17), Theorem 3.20 holds with the maximum replaced by C_n, thus we have

$$C_n - \sum_{t=1}^{n}\sum_{i=1}^{K}p_{i,t}v_{i,t} \leqslant \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \frac{\rho^2}{2}\sum_{t=1}^{n}A_t - \sum_{t=1}^{n}\sum_{i=1}^{K}v_{i,t}.$$

By taking the expectation we obtain

$$\begin{aligned}
&\mathbb{E}\left(C_n - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} v_{i,t}\right) \\
&\leqslant \frac{q}{q-1}\eta K^{1/q} + \gamma \mathbb{E} C_n + \mathbb{E}\sum_{t=1}^{n}\left(\frac{qc\mu^2(1+\mu^2)}{2\eta}\left(\sum_{i=1}^{K}\mathbb{E}_t v_{i,t}\right)^{(q-1)/q} - \sum_{i=1}^{K}\mathbb{E}_t v_{i,t}\right) \\
&\leqslant \frac{q}{q-1}\eta K^{1/q} + \gamma \mathbb{E} C_n + n\max_{u\geqslant 0}\left(\frac{qc\mu^2(1+\mu^2)}{2\eta}u^{\frac{q-1}{q}} - \frac{\gamma}{K}u\right) \\
&\leqslant \frac{q}{q-1}\eta K^{1/q} + \gamma \mathbb{E} C_n + \frac{\gamma n}{(q-1)K}\left(\frac{(q-1)cK\mu^2(1+\mu^2)}{2\gamma\eta}\right)^q.
\end{aligned}$$

The proof of (3.8) is concluded by using Inequality (3.2).

Proof of Theorem 3.9. We make use of (3.4). Since we have $B_t \leqslant Z_t/\varepsilon$ and $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$, we obtain

$$\left(\max_{1\leqslant i\leqslant K}\sum_{t=1}^{n} g_{i,t}\frac{Z_t}{\varepsilon}\right) - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} g_{i,t}\frac{Z_t}{\varepsilon} \leqslant \frac{\log K}{\eta} + \frac{\exp(2\eta/\varepsilon)\eta}{2\varepsilon^2}\sum_{t=1}^{n} Z_t,$$

hence, by taking the expectation of both sides,

$$\overline{R}_n = \left(\max_{1\leqslant i\leqslant K}\mathbb{E}\sum_{t=1}^{n} g_{i,t}\frac{Z_t}{\varepsilon}\right) - \mathbb{E}\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} g_{i,t}\frac{Z_t}{\varepsilon} \leqslant \frac{\log K}{\eta} + \frac{\exp(2\eta/\varepsilon)n\eta}{2\varepsilon}.$$

Straightforward computations conclude the proof.

Proof of Theorem 3.10. We start by noting that, since $R_n \leqslant n$, the result is trivial for $\delta \leqslant 2K\exp(-m/27)$ so that we consider hereafter that $\delta \geqslant 2K\exp(-m/27)$, or equivalently $\frac{\log(2K\delta^{-1})}{m} \leqslant \frac{1}{27}$.

From (10.3), we have

$$\mathbb{P}\left(\sum_{t=1}^{n} Z_t > m\right) \leqslant \exp\left(-\frac{m^2/16}{3m/2 + m/6}\right) \leqslant \exp\left(-\frac{m}{27}\right) \leqslant \frac{\delta}{4} \tag{3.19}$$

So with probability $1-\delta/4$, we have $\sum_{t=1}^{n} Z_t \leqslant m$, so that the rewards received by the forecaster are equal to the rewards which would receive the forecaster that uses Z_t to decide whether he asks for the gains or not, whatever $\sum_{s=1}^{t-1} Z_s$ is. This will enable us to use (3.4) (which holds with probability one).

From the concentration of martingales with bounded differences, see Theorem 10.1, with probability at least $1-\delta/4$, we also have

$$-\sum_{t=1}^{n} g_{I_t,t} \leqslant -\sum_{t=1}^{n}\sum_{k=1}^{K} p_{k,t} g_{k,t} + \sqrt{\frac{n\log(4\delta^{-1})}{2}}. \tag{3.20}$$

From Theorem 10.2, if $\eta\exp(2\eta\varepsilon^{-1}) \leqslant 2\varepsilon$ (which will be true for our particular η, see below), for a fixed $i \in \{1,\dots,K\}$, with probability at least $1-\delta/(2K)$, we have

$$\sum_{t=1}^{n}\left(g_{i,t} - \frac{\eta\exp(2\eta\varepsilon^{-1})}{2\varepsilon} - \sum_{k=1}^{K} p_{k,t} g_{k,t}\right)\left(1 - \frac{Z_t}{\varepsilon}\right) \leqslant 2\sqrt{\frac{2n\log(2K\delta^{-1})}{\varepsilon}} + \frac{2\log(2K\delta^{-1})}{3\varepsilon}.$$

From the union bound, we obtain that with probability at least $1-\delta/2$, we have

(3.21)
$$\max_{i=1,\dots,K}\sum_{t=1}^{n}\left(g_{i,t}-\frac{\eta\exp(2\eta\varepsilon^{-1})}{2\varepsilon}-\sum_{k=1}^{K}p_{k,t}g_{k,t}\right)\left(1-\frac{Z_t}{\varepsilon}\right)\leqslant 2\sqrt{\frac{2n\log(2K\delta^{-1})}{\varepsilon}}+\frac{2\log(2K\delta^{-1})}{3\varepsilon}$$

By combining (3.20) and (3.21), we obtain that with probability at least $1-3\delta/4$,

$$\begin{aligned}\max_{i=1,\dots,K}\sum_{t=1}^{n}g_{i,t}-\sum_{t=1}^{n}g_{I_t,t}&\leqslant\sqrt{\frac{n\log(4\delta^{-1})}{2}}+\max_{i=1,\dots,K}\sum_{t=1}^{n}g_{i,t}-\sum_{t=1}^{n}\sum_{k=1}^{K}p_{k,t}g_{k,t}\\&\leqslant\sqrt{\frac{n\log(4\delta^{-1})}{2}}+\max_{i=1,\dots,K}\sum_{t=1}^{n}g_{i,t}\frac{Z_t}{\varepsilon}-\sum_{t=1}^{n}\sum_{k=1}^{K}p_{k,t}g_{k,t}\frac{Z_t}{\varepsilon}\\&\quad+\sum_{t=1}^{n}\frac{\eta\exp(2\eta\varepsilon^{-1})}{2\varepsilon}\left(1-\frac{Z_t}{\varepsilon}\right)+2\sqrt{\frac{2n\log(2K\delta^{-1})}{\varepsilon}}+\frac{2\log(2K\delta^{-1})}{3\varepsilon}\end{aligned}$$

Now, by using (3.19) and (3.4), with probability at least $1-\delta$, we have

$$\begin{aligned}\max_{i=1,\dots,K}\sum_{t=1}^{n}g_{i,t}-\sum_{t=1}^{n}g_{I_t,t}&\leqslant\sqrt{\frac{n\log(4\delta^{-1})}{2}}+\frac{\log K}{\eta}+\frac{n\eta\exp(2\eta\varepsilon^{-1})}{2\varepsilon}\\&\quad+2n\sqrt{\frac{8\log(2K\delta^{-1})}{3m}}+\frac{8n\log(2K\delta^{-1})}{9m},\end{aligned}$$

hence, from the inequalities $m\leqslant n$, $K\geqslant 2$ and $\frac{\log(2K\delta^{-1})}{m}\leqslant\frac{1}{27}$,

$$\max_{i=1,\dots,K}\sum_{t=1}^{n}g_{i,t}-\sum_{t=1}^{n}g_{I_t,t}\leqslant n\sqrt{\frac{27\log(2K\delta^{-1})}{m}}+\frac{\log K}{\eta}+\frac{2\eta n^2\exp(8\eta n/(3m))}{3m}.$$

The inequality for $\eta=\frac{\sqrt{m\log K}}{n}$ is then obtained by noticing that the result needs to be proved only for $(2+\sqrt{27})\sqrt{(\log K)/m}<1$. This inequality is used to bound the exponential term (and in particular to check $\eta\exp(2\eta\varepsilon^{-1})\leqslant 2\varepsilon$). The second inequality is obtained by integrating the deviations using the standard formula $\mathbb{E}W\leqslant\int_0^1\frac{1}{\delta}\mathbb{P}(W>\log(\delta^{-1}))d\delta$ with $W=\frac{m}{27n^2}\left\{\max\left(0,R_n-2n\sqrt{\frac{\log K}{m}}\right)\right\}^2-\log(2K)$.

Proof of Theorem 3.11. The proof goes exactly like for Theorem 3.10. Using (3.7), one can prove that for any $q>1$, with probability at least $1-\delta$, INF satisfies:

$$R_n\leqslant\frac{q}{q-1}\eta K^{1/q}+\exp\left(\frac{16qn}{3m\eta}\right)\frac{2qn^2}{3m\eta}+n\sqrt{\frac{27\log(2K\delta^{-1})}{m}}.$$

Proof of Theorem 3.12. One simply need to note that for $\gamma\geqslant 3K\eta$, (3.5) is satisfied (since $B=K/\gamma$), and thus (3.6) rewrites

$$\max_{1\leqslant i\leqslant K}\sum_{t=1}^{n}\frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}-\sum_{t=1}^{n}g_{I_t,t}\leqslant\frac{1}{\eta}\log\left(\frac{K}{1-\gamma}\right)+\frac{\gamma}{1-\gamma}n.$$

By taking the expectation, one can see that the same bound holds for $\overline{R}_n$. The numerical application is proved by noticing that the bound is trivial for $\sqrt{20K\log(2K)}\geqslant\sqrt{n}$, whereas for $\sqrt{20K\log(2K)}<\sqrt{n}$, it uses $1-\gamma\geqslant 1-(20)^{-1/2}$ and straightforward computations.

Proof of Theorem 3.13. The inequality is trivial when $\sqrt{K/n} \geqslant 1/10$. So we now consider that $\sqrt{K/n} < 1/10$ and apply (3.8). For the chosen values of q, η and γ, we have $\frac{1}{\eta}(\frac{K}{\gamma})^{(q-1)/q} = \frac{1}{3}(\frac{K}{n})^{1/4} \leqslant \frac{1}{3\sqrt{10}}$, hence $\mu \leqslant 2.1$, and, $\overline{R}_n \leqslant 9/0.9\sqrt{nK}$.

Proof of Theorem 3.14. Note that for $\gamma \geqslant 3K\eta(1+2\beta)$, (3.5) is satisfied (since $B = (1+\beta)K/\gamma$), and thus (3.6) rewrites

$$\max_{1\leqslant i\leqslant K} \sum_{t=1}^{n} \frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{p_{i,t}} - \sum_{t=1}^{n} g_{I_t,t} - \beta nK \leqslant \frac{\gamma}{1-\gamma}(n + \beta nK) + \frac{1}{\eta}\log\left(\frac{K}{1-\gamma}\right).$$

Let $\mathbb{E}_t$ be the expectation resulting from $I_t \sim p_t$. Since $\exp(x) \leqslant 1 + x + x^2$ for $x \leqslant 1$, we have for $\beta \leqslant 1$ and a fixed i

$$\begin{aligned}
&\mathbb{E}_t \exp\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{p_{i,t}}\right)\\
\leqslant&\left\{1 + \mathbb{E}_t\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\right) + \mathbb{E}_t\left(\beta g_{i,t} - \beta\frac{g_{i,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\right)^2\right\}\exp\left(-\frac{\beta^2}{p_{i,t}}\right)\\
\leqslant&\left\{1 + \beta^2\frac{g_{i,t}^2}{p_{i,t}}\right\}\exp\left(-\frac{\beta^2}{p_{i,t}}\right)\\
\leqslant&1,
\end{aligned}$$

where the last inequality uses $1 + u \leqslant \exp(u)$. As a consequence, we have

$$\mathbb{E}\exp\left(\beta\sum_{t=1}^{n} g_{i,t} - \beta\sum_{t=1}^{n}\frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{p_{i,t}}\right) \leqslant 1.$$

Moreover Markov's inequality implies $\mathbb{P}\left(X > \log(\delta^{-1})\right) \leqslant \delta\mathbb{E}e^X$ and thus with probability at least $1 - \delta/K$

$$\beta\sum_{t=1}^{n} g_{i,t} - \beta\sum_{t=1}^{n}\frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{p_{i,t}} \leqslant \log(K\delta^{-1}). \tag{3.22}$$

From a union bound, this implies that with probability at least $1 - \delta$, we have

$$\max_{1\leqslant i\leqslant K}\left(\sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n}\frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{p_{i,t}}\right) \leqslant \frac{\log(K\delta^{-1})}{\beta}. \tag{3.23}$$

Thus we get, with probability at least $1 - \delta$,

$$R_n \leqslant \frac{(\gamma + \beta K)n}{1-\gamma} + \frac{1}{\eta}\log\left(\frac{K}{1-\gamma}\right) + \frac{1}{\beta}\log\left(\frac{K}{\delta}\right).$$

The numerical application is proved by noticing that the bound is trivial for $\sqrt{K\log(2K)/n} \geqslant 1/7$, whereas for $\sqrt{K\log(2K)/n} < 1/7$, it uses $1 - \gamma \geqslant 1 - \sqrt{2}/7$ and straightforward computations. The inequality on $\mathbb{E}R_n$ is then obtained by integrating the deviations using $\mathbb{E}W \leqslant \int_0^1 \frac{1}{\delta}\mathbb{P}(W > \log(\delta^{-1}))d\delta$ for W a real-valued random variable.

Proof of Theorem 3.15. Actually we prove the following result. For $\eta > 0$, $q > 1$ and $\gamma \in (0,1)$ such that $q\eta > (1+\beta)\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}$. Introduce

$$\widetilde{\mu} = \exp\left\{\frac{2(q+1)K}{\eta\gamma}\left[\left(\frac{\gamma}{K}\right)^{1/q}\left(1 - \frac{1+\beta}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q} + \beta\right]\right\}$$

and

$$\zeta = \frac{q\widetilde{\mu}(1+\widetilde{\mu})(1+2\beta)}{2\eta}.$$

We assume $\zeta^q K^{q-1} \leqslant \gamma^{q-1}$. Then in the bandit game, for any $\delta > 0$ and any $i \in \{1, \dots, K\}$, with probability at least $1-\delta$, INF satisfies:

$$\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{I_t,t} \leqslant \frac{1}{1-\gamma}\left\{(\gamma+\beta K)n + \frac{q\eta K^{\frac{1}{q}}}{q-1} + \frac{\gamma n(1+\beta^{q+1}K)}{(q-1)K}\left(\frac{(q-1)K\zeta}{q\gamma}\right)^q\right\}$$
$$+ \frac{1}{1-\gamma}\left\{\sqrt{2n\max(\gamma, K\zeta^2)\log(2\delta^{-1})} + \frac{2\log(2\delta^{-1})}{3}\right\} + \frac{\log(2\delta^{-1})}{\beta}. \tag{3.24}$$

Besides, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:

$$R_n \leqslant \frac{1}{1-\gamma}\left\{(\gamma+\beta K)n + \frac{q\eta K^{\frac{1}{q}}}{q-1} + \frac{\gamma n(1+\beta^{q+1}K)}{(q-1)K}\left(\frac{(q-1)K\zeta}{q\gamma}\right)^q\right\}$$
$$+ \frac{1}{1-\gamma}\left\{\sqrt{2n\max(\gamma, K\zeta^2)\log(2\delta^{-1})} + \frac{2\log(2\delta^{-1})}{3}\right\} + \frac{\log(2K\delta^{-1})}{\beta}, \tag{3.25}$$

The starting point of the proof is similar to the one of Theorem 3.6. We make use of Theorem 3.20 (and its notations) and straightforward computations to obtain, since $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i} + \frac{\beta}{p_{i,t}}$,

$$C_n - \sum_{t=1}^n g_{I_t,t} - \beta nK \leqslant \frac{q}{q-1}\eta K^{1/q} + \gamma\left(C_n - \frac{1}{K}\sum_{t=1}^n\sum_{i=1}^K \frac{g_{i,t}\mathbb{1}_{I_t=i}+\beta}{p_{i,t}}\right)$$
$$+ \frac{q\rho^2(1+\rho^2)}{2\eta}\sum_{t=1}^n\sum_{i=1}^K \frac{g_{i,t}(1+2\beta)\mathbb{1}_{I_t=i}+\beta^2}{p_{i,t}^{(q-1)/q}}, \tag{3.26}$$

where we used $(g_{i,t}\mathbb{1}_{I_t=i}+\beta)^2 \leqslant g_{i,t}(1+2\beta)\mathbb{1}_{I_t=i}+\beta^2$ for the last equation. Now note that for any $a \geqslant b > c > 0$ we have $\frac{a}{b} \leqslant \frac{a-c}{b-c}$ and thus for any $x < 0$, we have

$$\frac{\psi\left(x+\frac{1+\beta}{\psi(x)}\right)}{\psi(x)} \leqslant \frac{\psi\left(x+\frac{1+\beta}{\psi(x)}\right)-\frac{\gamma}{K}}{\psi(x)-\frac{\gamma}{K}} = \left(1-\frac{1+\beta}{-x\psi(x)}\right)^{-q} \leqslant \left(1-\frac{1+\beta}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{\frac{q-1}{q}}\right)^{-q}.$$

Besides we have $\frac{\psi''}{\psi'} = \frac{q+1}{\eta}(\psi-\gamma/K)^{1/q} \leqslant \frac{q+1}{\eta}\psi^{1/q}$, and thus lemma 3.4 gives us

$$\rho^2 \leqslant \exp\left\{\frac{2(q+1)K}{\eta\gamma}\left[\left(\frac{\gamma}{K}\right)^{1/q}\left(1-\frac{1+\beta}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}+\beta\right]\right\} = \widetilde{\mu}. \tag{3.27}$$

Let us introduce $\zeta = \frac{q\widetilde{\mu}(1+\widetilde{\mu})(1+2\beta)}{2\eta}$, and $G_{k,n} = \sum_{t=1}^n g_{k,t}$ for a fixed $k \in \{1, \dots, K\}$,

$$D_1 = C_n - G_{k,n},$$

$$X_t = \sum_{i=1}^K \left(\zeta g_{i,t} p_{i,t}^{1/q} - \frac{\gamma}{K} g_{i,t}\right)\left(\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} - 1\right),$$

$$D_2 = \sum_{t=1}^n X_t,$$

$$S_1 = \sum_{t=1}^n\sum_{i=1}^K \left(\zeta g_{i,t} p_{i,t}^{1/q} - \frac{\gamma}{K} g_{i,t}\right),$$

and

$$S_2 = \sum_{t=1}^{n}\sum_{i=1}^{K}\left(\zeta\frac{\beta^2}{1+2\beta}p_{i,t}^{1/q} - \frac{\gamma}{K}\beta\right)\frac{1}{p_{i,t}}.$$

Plugging (3.27) into (3.26), we obtain

$$(1-\gamma)\Big(D_1 + G_{k,n}\Big) - \sum_{t=1}^{n} g_{I_t,t} \leqslant \beta n K + \frac{q}{q-1}\eta K^{1/q} + D_2 + S_1 + S_2. \tag{3.28}$$

We will now provide high probability bounds for D_1 and D_2, and upper bounds with probability one for S_1 and S_2. Hereafter, we consider a confidence level $\delta > 0$.

Lower bound on D_1: By using Inequality (3.2), we have $D_1 \geqslant \sum_{t=1}^{n}\frac{g_{k,t}\mathbb{1}_{I_t=k}+\beta}{p_{k,t}} - G_{k,n}$. From the argument which lead to (3.22), we have with probability at least $1-\delta/2$, we have

$$\beta G_{k,n} - \beta\sum_{t=1}^{n}\frac{g_{k,t}\mathbb{1}_{I_t=k}+\beta}{p_{k,t}} \leqslant \log(2\delta^{-1}), \tag{3.29}$$

hence

$$D_1 \geqslant -\frac{\log(2\delta^{-1})}{\beta}.$$

Upper bound on D_2: To apply the Bernstein inequality given in Theorem 10.2, we need to upper bound $|X_t|$ and $\mathbb{E}_t X_t^2$, where $\mathbb{E}_t$ still denotes the expectation resulting from $I_t \sim p_t$. Since we have

$$X_t = \frac{\zeta g_{I_t,t}p_{I_t,t}^{1/q} - \frac{\gamma}{K}g_{I_t,t}}{p_{I_t,t}} - \sum_{i=1}^{K}\left(\zeta g_{i,t}p_{i,t}^{1/q} - \frac{\gamma}{K}g_{i,t}\right), \tag{3.30}$$

we have $X_t \leqslant \zeta\big(\frac{K}{\gamma}\big)^{(q-1)/q} + \gamma \leqslant 1+\gamma$ and $X_t \geqslant -1 - \zeta\sum_{i=1}^{K}p_{i,t}^{1/q} \geqslant -1 - \zeta K^{(q-1)/q} \geqslant -2$, where we have used Hölder's inequality and $\sum_{i=1}^{K}p_{i,t}=1$. Therefore, we have $|X_t| \leqslant 2$.

From (3.30) and the inequalities $\gamma < 1 \leqslant K$, $\zeta K^{(q-1)/q} \leqslant \gamma^{(q-1)/q}$, we have

$$\begin{aligned}
\mathbb{E}X_t^2 = \mathbb{V}\mathrm{ar}X_t &\leqslant \mathbb{E}\left(\frac{\zeta g_{I_t,t}p_{I_t,t}^{1/q} - \frac{\gamma}{K}g_{I_t,t}}{p_{I_t,t}}\right)^2 \\
&\leqslant \sum_{i=1}^{K}\frac{\big(\zeta p_{i,t}^{1/q} - \frac{\gamma}{K}\big)^2}{p_{i,t}} \\
&\leqslant K\max_{\frac{\gamma}{K}\leqslant u\leqslant 1}\frac{\big(\zeta u^{1/q} - \frac{\gamma}{K}\big)^2}{u} \\
&= K\max\left\{\frac{K}{\gamma}\left(\zeta\Big(\frac{\gamma}{K}\Big)^{1/q} - \frac{\gamma}{K}\right)^2, \Big(\zeta - \frac{\gamma}{K}\Big)^2\right\} = K\max\Big\{\frac{\gamma}{K}, \zeta^2\Big\}.
\end{aligned}$$

Theorem 10.2 implies that with probability at least $1-\delta/2$, we have

$$D_2 = \sum_{t=1}^{n}X_t \leqslant \sqrt{2n\max(\gamma, K\zeta^2)\log(2\delta^{-1})} + \frac{2\log(2\delta^{-1})}{3}.$$

Upper bound on S_1: Following the same arguments as in the proof of (3.8), that is essentially Hölder's inequality followed by an optimization with respect to the value of $\sum_{i=1}^{n}g_{i,t}$, we have

$$S_1 = \sum_{t=1}^{n}\sum_{i=1}^{K}\left(\zeta g_{i,t}p_{i,t}^{1/q} - \frac{\gamma}{K}g_{i,t}\right) \leqslant n\max_{u\geqslant 0}\left(\zeta u^{\frac{q-1}{q}} - \frac{\gamma}{K}u\right) = \frac{\gamma n}{(q-1)K}\left(\frac{(q-1)K\zeta}{q\gamma}\right)^q.$$

Upper bound on S_2: We have

$$S_2 \leqslant \beta n K \max_{u>0} \left(\zeta \frac{\beta}{1+2\beta} u^{\frac{q-1}{q}} - \frac{\gamma}{K} u \right) = \frac{\gamma\beta n}{q-1} \left(\frac{(q-1)K\zeta\beta}{(1+2\beta)q\gamma} \right)^q \leqslant \frac{\gamma\beta^{q+1} n}{q-1} \left(\frac{(q-1)K\zeta}{q\gamma} \right)^q .$$

Putting the previous bounds into (3.28), we get the first inequality of the theorem. The second inequality is obtained by replacing (3.29) by

$$\beta \max_{k=1,\dots,K} G_{k,n} \leqslant \beta \max_{k=1,\dots,K} \sum_{t=1}^{n} \frac{g_{k,t} \mathbb{1}_{I_t=k} + \beta}{p_{k,t}} + \log(2K\delta^{-1}) \leqslant \beta C_n + \log(2K\delta^{-1}),$$

which, from a union bound, also holds with probability at least $1 - \delta/2$.

The numerical application (3.12) is proved by using $2\sqrt{\log(2\delta^{-1})} \leqslant 1 + \log(2\delta^{-1})$ as well as noticing that the bound is trivial for $10.7\sqrt{nK} \geqslant n$, whereas for $10.7\sqrt{nK} < n$, one can successively check that $\gamma = \sqrt{K/n} \leqslant 1/10.7$, $\beta = \sqrt{\log(2K)/nK} \leqslant 1/4.2$, $\widetilde{\mu} \leqslant 2.92$ and $\zeta \leqslant 5.63/\sqrt{n}$, which leads to the desired result after straightforward computations. The same reasonning leads to (3.11). The inequalities on $\mathbb{E}R_n$ are then obtained by integrating the deviations using $\mathbb{E}W \leqslant \int_0^1 \frac{1}{\delta} \mathbb{P}(W > \log(\delta^{-1}))d\delta$ for W a real-valued random variable.

Proofs of Theorem 3.16 and Theorem 3.17. The proof of Theorem 3.16 goes exactly like for Theorem 3.12, for $\gamma\varepsilon \geqslant 3K\eta$ we have

$$\overline{R}_n \leqslant \frac{1}{\eta} \log \left(\frac{K}{1-\gamma} \right) + \frac{\gamma}{1-\gamma} n.$$

The proof of Theorem 3.17 is also trivial.

Proof of Theorem 3.18. We prove that for any $\eta > 0$ and $\gamma \in (0,1)$ such that $\gamma \geqslant 3K\eta(1+2\beta)$ and $\gamma\widetilde{\rho} < K\beta$ with $\widetilde{\rho} = \frac{1+K\beta}{1-\gamma} \exp\left((1+\beta)\frac{K\eta}{\gamma}\right)$, INF satisfies:

$$R_n^S \leqslant \frac{(\gamma+\beta K)n}{1-\gamma} + \frac{1}{\eta} \log \left(\frac{K}{1-\gamma} \right) + \frac{1}{\beta} \log \left(\frac{K}{\delta} \right) + S \left\{ \frac{1}{\beta} \log \left(\frac{enK}{S} \right) - \frac{1}{\eta} \log \left(\frac{\beta}{\widetilde{\rho}} - \frac{\gamma}{K} \right) + \widetilde{\rho} \right\},$$

First note that, as we have already seen in the proof of Theorem 3.14, (3.6) gives

$$\left(\max_{1 \leqslant i \leqslant K} V_{i,n} \right) - \sum_{t=1}^{n} g_{I_t,t} \leqslant \frac{(\gamma+\beta K)n}{1-\gamma} + \frac{1}{\eta} \log \left(\frac{K}{1-\gamma} \right).$$

Let $\xi_t = \max_{i=1,\dots,K} V_{i,t} - \min_{j=1,\dots,K} V_{j,t}$ and $\xi = \max_{t=1,\dots,n} \xi_t$. Consider a fixed switching strategy $(i_1, \dots, i_n) \in \{1, \dots, K\}^n$, and let $V_{(i_1,\dots,i_n)} = \sum_{t=1}^{n} v_{i_t,t}$. One can easily check that

$$\max_{1 \leqslant i \leqslant K} V_{i,n} \geqslant V_{(i_1,\dots,i_n)} - \xi \mathcal{S}(i_1, \dots, i_n).$$

Let $M = \sum_{j=0}^{S} \binom{n-1}{j} K(K-1)^j$ be the number of switching strategies of size not larger than S. The argument which leads to (3.22) can be used to prove that with probability at least $1 - \delta/M$, we have

$$\beta G_{(i_1,\dots,i_n)} - \beta V_{(i_1,\dots,i_n)} \leqslant \log(M\delta^{-1}).$$

By putting the three previous inequalities together, we obtain that with probability at least $1 - \delta/M$,

$$G_{(i_1,\dots,i_n)} - \sum_{t=1}^{n} g_{I_t,t} \leqslant \frac{(\gamma+\beta K)n}{1-\gamma} + \frac{1}{\eta} \log \left(\frac{K}{1-\gamma} \right) + \xi \mathcal{S}(i_1, \dots, i_n) + \frac{1}{\beta} \log \left(\frac{M}{\delta} \right).$$

From a union bound, with probability at least $1-\delta$, we have

$$\max_{(i_1,\dots,i_n)\colon \mathcal{S}(i_1,\dots,i_n)\leqslant S} G_{(i_1,\dots,i_n)} - \sum_{t=1}^n g_{I_t,t} \leqslant \frac{(\gamma+\beta K)n}{1-\gamma} + \frac{1}{\eta}\log\left(\frac{K}{1-\gamma}\right) + \frac{1}{\beta}\log\left(\frac{M}{\delta}\right) + S\xi,$$

which is the desired result up to appropriate upper bounds on M and ξ. We have

$$M = \sum_{j=0}^S \binom{n-1}{j} K(K-1)^j \leqslant K^{S+1}\sum_{j=0}^S \binom{n-1}{j} \leqslant K^{S+1}\left(\frac{en}{S}\right)^S,$$

where the second inequality comes from Sauer's lemma. Now, by contradiction, we will prove

$$\xi \leqslant \widetilde{\rho} - \frac{1}{\eta}\log\left(\frac{\beta}{\widetilde{\rho}} - \frac{\gamma}{K}\right). \tag{3.31}$$

To this end, we start by bounding $C_t - C_{t-1}$. By the mean value theorem, with the notations of the third step of the proof of Theorem 3.20, there exists $W \in [V_{t-1}, V_t]$ such that

$$\begin{aligned}
C_t - C_{t-1} &= C(V_t) - C(V_{t-1}) \\
&= \sum_{i=1}^K \frac{\partial C}{\partial x_i}(W)(V_{i,t} - V_{i,t-1}) \\
&= \sum_{i=1}^K \frac{h_i(W)}{\sum_{j=1}^K h_j(W)} \frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{f_i(V_{i,t-1})} \\
&= \frac{1}{\sum_{j=1}^K \eta(f_j(W) - \gamma/K)} \sum_{i=1}^K \eta h_i(W) \frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{h_i(V_{i,t-1}) + \eta\gamma/K} \\
&\leqslant \frac{1}{1-\gamma}\sum_{i=1}^K h_i(W)\frac{\mathbb{1}_{I_t=i} + \beta}{h_i(V_{t-1})} \leqslant \frac{\rho}{1-\gamma}\sum_{i=1}^K (\mathbb{1}_{I_t=i} + \beta) = \rho\frac{1+K\beta}{1-\gamma}.
\end{aligned}$$

From Lemma 3.2, we have $\rho \leqslant \exp\left((1+\beta)\frac{K\eta}{\gamma}\right)$, hence $C_t - C_{t-1} \leqslant \exp\left((1+\beta)\frac{K\eta}{\gamma}\right)\frac{1+K\beta}{1-\gamma} = \widetilde{\rho}$. If (3.31) does not hold, then from Lemma 3.1, there exist $T \in \{1,\dots,n\}$ and $\ell \in \{1,\dots,K\}$ such that $C_{T-1} - V_{\ell,T-1} \leqslant \widetilde{\rho} - \psi^{-1}(\beta/\widetilde{\rho})$ and $C_T - V_{\ell,T} > \widetilde{\rho} - \psi^{-1}(\beta/\widetilde{\rho})$ (note that we have $C_0 - V_{\ell,0} = -\psi^{-1}(1/K) \leqslant \widetilde{\rho} - \psi^{-1}(\beta/\widetilde{\rho})$ since $K\beta \leqslant \widetilde{\rho}$). In particular, we have $\psi(V_{\ell,T} - C_T + \widetilde{\rho}) < \frac{\beta}{\widetilde{\rho}}$, hence

$$V_{\ell,T} - V_{\ell,T-1} \geqslant \frac{\beta}{p_{\ell,T}} = \frac{\beta}{\psi(V_{\ell,T-1} - C_{T-1})} \geqslant \frac{\beta}{\psi(V_{\ell,T} - C_T + \widetilde{\rho})} \geqslant \widetilde{\rho} \geqslant C_T - C_{T-1},$$

which contradicts the inequality $C_{T-1} - V_{\ell,T-1} < C_T - V_{\ell,T}$. This ends the proof of (3.31), and consequently of the first inequality of the theorem.

The numerical application given in the second inequality is proved by noticing that the bound is trivial for $9\sqrt{Ks} \geqslant \sqrt{n}$, whereas for $9\sqrt{Ks} < \sqrt{n}$, it uses $1-\gamma \geqslant 8/9$, $\widetilde{\rho} \leqslant \sqrt{3}$, $\beta \leqslant \frac{2}{9}$, $\frac{1}{\eta}\log\left(\frac{K}{1-\gamma}\right) \leqslant \sqrt{20}\log(2K)$, $\frac{1}{\beta(1-\gamma)}\log\left(\frac{K}{\delta}\right) \leqslant \frac{9}{16}\log(K\delta^{-1})$, $\frac{S\widetilde{\rho}}{1-\gamma} \leqslant \frac{S}{9}\sqrt{nK/s}$, $-\frac{1}{\eta}\log\left(\frac{\beta}{\widetilde{\rho}} - \frac{\gamma}{K}\right) \leqslant 3S\log\left(\frac{enK}{S}\right)\sqrt{\frac{nK}{s}}$ and straightforward computations. The last inequality follows by integrating the deviations.

Proof of Theorem 3.19. We may assume $\mu_1 \geqslant \ldots \geqslant \mu_K$. Using the trivial equality $\sum_{i=1}^K \mathbb{E}T_i(n) = n$, we have

$$\overline{R}_n = \max_{i=1,\dots,K} \mathbb{E}\sum_{t=1}^n \left(g_{i,t} - g_{I_t,t}\right)$$

$$
\begin{aligned}
&= n\Big(\max_{i=1,\dots,K} \mathbb{E}g_{i,t}\Big) - \sum_{t=1}^{n} \mathbb{E}g_{I_t,t} \\
&= n\Big(\max_{i=1,\dots,K} \mu_i\Big) - \sum_{t=1}^{n} \mathbb{E}\mu_{I_t} \\
&= n\Big(\max_{i=1,\dots,K} \mu_i\Big) - \mathbb{E}\sum_{t=1}^{n} \mu_{I_t} \\
&= \Big(\sum_{i=1}^{K} \mathbb{E}T_i(n)\Big)\Big(\max_{i=1,\dots,K} \mu_i\Big) - \mathbb{E}\sum_{i=1}^{K} \mu_i T_i(n) = \sum_{i=1}^{K} \Delta_i \mathbb{E}T_i(n).
\end{aligned}
$$

First step: Decoupling the arms. For an arm k_0, we trivially have $\sum_{k=1}^{K} \Delta_k T_k(n) \leqslant n\Delta_{k_0} + \sum_{k=k_0+1}^{K} \Delta_k T_k(n)$. Let $\Delta_{K+1} = +\infty$, $z_k = \mu_1 - \frac{\Delta_k}{2}$ for $k_0 < k \leqslant K+1$ and $z_{k_0} = +\infty$. Let $Z = \min_{1\leqslant s\leqslant n} B_{1,s}$ and $W_{j,k} = \mathbb{1}_{Z\in[z_{j+1},z_j)}(\Delta_k - \Delta_{k_0})T_k(n)$. By using $\mathbb{E}\sum_{k=1}^{k_0} T_k(n) = n - \mathbb{E}\sum_{k=k_0+1}^{K} T_k(n)$, we get

$$
\overline{R}_n = \mathbb{E}\sum_{k=1}^{K} \Delta_k T_k(n) \leqslant n\Delta_{k_0} + \mathbb{E}\sum_{k=k_0+1}^{K} (\Delta_k - \Delta_{k_0})T_k(n).
$$

We have

$$
\sum_{k=k_0+1}^{K} (\Delta_k - \Delta_{k_0})T_k(n) = \sum_{k=k_0+1}^{K}\sum_{j=k_0}^{K} W_{j,k} = \sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j} W_{j,k} + \sum_{j=k_0}^{K}\sum_{k=j+1}^{K} W_{j,k}. \tag{3.32}
$$

An Abel transformation takes care of the first sum of (3.32):

$$
\sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j} W_{j,k} \leqslant \sum_{j=k_0}^{K} \mathbb{1}_{Z\in[z_{j+1},z_j)} n(\Delta_j - \Delta_{k_0}) = n\sum_{j=k_0+1}^{K} \mathbb{1}_{Z<z_j}(\Delta_j - \Delta_{j-1}). \tag{3.33}
$$

To bound the second sum of (3.32), we introduce the stopping times $\tau_k = \min\{t : B_{k,t} < z_k\}$ and remark that, by definition of MOSS, we have $\{Z \geqslant z_k\} \subset \{T_k(n) \leqslant \tau_k\}$, since once we have pulled τ_k times arm k its index will always be lower than the index of arm 1. This implies

$$
\sum_{j=k_0}^{K}\sum_{k=j+1}^{K} W_{j,k} = \sum_{k=k_0+1}^{K}\sum_{j=k_0}^{k-1} W_{j,k} = \sum_{k=k_0+1}^{K} \mathbb{1}_{Z\geqslant z_k}\Delta_k T_k(n) \leqslant \sum_{k=k_0+1}^{K} \tau_k\Delta_k. \tag{3.34}
$$

Combining (3.32), (3.33) and (3.34) and taking the expectation, we get

$$
\overline{R}_n \leqslant n\Delta_{k_0} + \sum_{k=k_0+1}^{K} \Delta_k \mathbb{E}\tau_k + n\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}). \tag{3.35}
$$

Let $\delta_0 = \sqrt{\frac{75K}{n}}$ and set k_0 such that $\Delta_{k_0} \leqslant \delta_0 < \Delta_{k_0+1}$. If $k_0 = K$, we trivially have $\overline{R}_n \leqslant n\delta_0 \leqslant \sqrt{75nK}$ so that (3.13) holds trivially. In the following, we thus consider $k_0 < K$.

Second step: Bounding $\mathbb{E}\tau_k$ for $k_0 + 1 \leqslant k \leqslant K$.

Let $\log_+(x) = \max(\log(x), 0)$. For $\ell_0 \in \mathbb{N}$, we have

$$
\mathbb{E}\tau_k - \ell_0 = \sum_{\ell=0}^{+\infty} \mathbb{P}(\tau_k > \ell) - \ell_0 \tag{3.36}
$$

$$\leqslant \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}(\tau_k > \ell) = \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}(\forall t \leqslant \ell, B_{k,t} > z_k)$$
$$\leqslant \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}\left(\widehat{\mu}_{k,\ell} - \mu_k \geqslant \frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/K\ell)}{\ell}}\right).$$

Now let us take $\ell_0 = \lceil 7\log\left(\frac{n}{K}\Delta_k^2\right)/\Delta_k^2 \rceil$ with $\lceil x \rceil$ the smallest integer larger than x. For $\ell \geqslant \ell_0$, since $k > k_0$, we have

$$\log_+\left(\frac{n}{K\ell}\right) \leqslant \log_+\left(\frac{n}{K\ell_0}\right) \leqslant \log_+\left(\frac{n\Delta_k^2}{7K}\right) \leqslant \frac{\ell_0\Delta_k^2}{7} \leqslant \frac{\ell\Delta_k^2}{7},$$

hence $\frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/(K\ell))}{\ell}} \geqslant c\Delta_k$, with $c = \frac{1}{2} - \frac{1}{\sqrt{7}}$. Therefore, by using Hoeffding's inequality and (3.36), we get

$$\mathbb{E}\tau_k - \ell_0 \leqslant \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}\left(\widehat{\mu}_{k,\ell} - \mu_k \geqslant c\Delta_k\right)$$
$$\leqslant \sum_{\ell=\ell_0}^{+\infty} \exp\left(-2\ell(c\Delta_k)^2\right) = \frac{\exp\left(-2\ell_0(c\Delta_k)^2\right)}{1-\exp\left(-2(c\Delta_k)^2\right)} \leqslant \frac{\exp(-14c^2\log(75))}{1-\exp\left(-2c^2\Delta_k^2\right)}, \tag{3.37}$$

where the last inequality uses $\ell_0\Delta_k^2 \geqslant 7\log(75)$. Plugging the value of ℓ_0, we obtain

$$\Delta_k\mathbb{E}\tau_k \leqslant \Delta_k\left(1 + \frac{7\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k^2}\right) + \frac{\Delta_k\exp(-14c^2\log(75))}{1-\exp\left(-2c^2\Delta_k^2\right)}$$
$$\leqslant 1 + 7\frac{\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k} + \frac{\exp(-14c^2\log(75))}{2c^2(1-c^2)\Delta_k}, \tag{3.38}$$

where the last step uses that, since $1 - \exp(-x) \geqslant x - x^2/2$ for any $x \geqslant 0$, we have

$$\frac{1}{1-\exp\left(-2c^2\Delta_k^2\right)} \leqslant \frac{1}{2c^2\Delta_k^2 - 2c^4\Delta_k^4} \leqslant \frac{1}{2c^2\Delta_k^2(1-c^2)}$$

Third step: Bounding $n\sum_{k=k_0+1}^{K}\mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1})$.

Let X_t denote the reward obtained by arm 1 when it is drawn for the t-th time. The random variables $X_1, X_2, \ldots$ are i.i.d. so that we have the maximal inequality [Hoeffding, 1963, Inequality (2.17)]: for any $x > 0$ and $m \geqslant 1$,

$$\mathbb{P}\left(\exists s \in \{1,\ldots,m\}, \sum_{t=1}^{s}(\mu_1 - X_t) > x\right) \leqslant \exp\left(-\frac{2x^2}{m}\right).$$

Since $z_k = \mu_1 - \Delta_k/2$ and since $u \mapsto \mathbb{P}\left(Z < \mu_1 - u/2\right)$ is a nonincreasing function, we have

$$\sum_{k=k_0+1}^{K}\mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}) \leqslant \Delta_{k_0+1}\mathbb{P}(Z < z_{k_0+1}) + \int_{\Delta_{k_0+1}}^{1}\mathbb{P}\left(Z < \mu_1 - \frac{u}{2}\right)du. \tag{3.39}$$

We will now concentrate on upper bounding $\mathbb{P}\left(Z < \mu_1 - \frac{u}{2}\right)$ for a fixed $u \in [\delta_0, 1]$. Let $f(u) = 8\log\left(\sqrt{\frac{n}{K}}u\right)/u^2$. We have

$$\mathbb{P}\left(Z < \mu_1 - \frac{1}{2}u\right) \quad = \quad \mathbb{P}\left(\exists 1 \leqslant s \leqslant n : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s\log_+\left(\frac{n}{Ks}\right)} + \frac{su}{2}\right)$$

$$\begin{aligned} \leqslant \quad & \mathbb{P}\left(\exists 1 \leqslant s \leqslant f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s \log_+\left(\frac{n}{Ks}\right)}\right) \\ + \quad & \mathbb{P}\left(\exists f(u) < s \leqslant n : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\right). \end{aligned}$$

For the first term, we use a peeling argument with a geometric grid of the form $\frac{1}{2^{\ell+1}} f(u) \leqslant s \leqslant \frac{1}{2^\ell} f(u)$. The numerical constant in δ_0 ensures that $f(u) \leqslant n/K$, which implies that for any $s \leqslant f(u)$, $\log_+\left(\frac{n}{Ks}\right) = \log\left(\frac{n}{Ks}\right)$. We have

$$\begin{aligned} & \mathbb{P}\left(\exists 1 \leqslant s \leqslant f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s \log\left(\frac{n}{Ks}\right)}\right) \\ & \leqslant \sum_{\ell=0}^{+\infty} \mathbb{P}\left(\exists \frac{1}{2^{\ell+1}} f(u) \leqslant s \leqslant \frac{1}{2^\ell} f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{\frac{f(u)}{2^{\ell+1}} \log\left(\frac{n2^\ell}{Kf(u)}\right)}\right) \\ & \leqslant \sum_{\ell=0}^{+\infty} \exp\left(-2\frac{f(u)\frac{1}{2^{\ell+1}} \log\left(\frac{n2^\ell}{Kf(u)}\right)}{f(u)\frac{1}{2^\ell}}\right) = \sum_{\ell=0}^{+\infty} \frac{Kf(u)}{n}\frac{1}{2^\ell} = \frac{16K}{nu^2} \log\left(\sqrt{\frac{n}{K}}u\right). \end{aligned}$$

For the second term we also use a peeling argument but with a geometric grid of the form $2^\ell f(u) \leqslant s \leqslant 2^{\ell+1} f(u)$:

$$\begin{aligned} & \mathbb{P}\Big(\exists s \in \lceil f(u) \rceil, \dots, n\} : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\Big) \\ & \qquad \leqslant \sum_{\ell=0}^{+\infty} \mathbb{P}\left(\exists 2^\ell f(u) \leqslant s \leqslant 2^{\ell+1} f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > 2^{\ell-1} f(u) u\right) \\ & \qquad \leqslant \sum_{\ell=0}^{+\infty} \exp\left(-2\frac{\left(2^{\ell-1} f(u) u\right)^2}{f(u) 2^{\ell+1}}\right) \\ & \qquad = \sum_{\ell=0}^{+\infty} \exp\left(-2^\ell f(u) u^2/4\right) \\ & \qquad \leqslant \sum_{\ell=0}^{+\infty} \exp\left(-(\ell+1) f(u) u^2/4\right) = \frac{1}{\exp\left(f(u)u^2/4\right) - 1} = \frac{1}{nu^2/K - 1}. \end{aligned}$$

Putting together the last three computations, we obtain

$$\mathbb{P}\left(Z < \mu_1 - \frac{1}{2}u\right) \leqslant \frac{16K}{nu^2} \log\left(\sqrt{\frac{n}{K}}u\right) + \frac{1}{nu^2/K - 1}.$$

Plugging this into (3.39) gives

$$\begin{aligned} & \sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}) \\ & \qquad \leqslant \frac{16K}{n\Delta_{k_0+1}} \log\left(\sqrt{\frac{n}{K}}\Delta_{k_0+1}\right) + \frac{\Delta_{k_0+1}}{n\Delta_{k_0+1}^2/K - 1} \\ & \qquad\quad + \left[-\frac{16K}{nu} \log\left(e\sqrt{\frac{n}{K}}u\right) + \sqrt{\frac{K}{4n}} \log\left(\frac{\sqrt{\frac{n}{K}}u - 1}{\sqrt{\frac{n}{K}}u + 1}\right)\right]_{\Delta_{k_0+1}}^{1} \end{aligned}$$

$$\leqslant \frac{16K}{n\Delta_{k_0+1}} \log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \frac{\Delta_{k_0+1}}{n\Delta_{k_0+1}^2/K - 1} + \sqrt{\frac{K}{4n}} \log\left(\frac{\sqrt{\frac{n}{K}}\Delta_{k_0+1} + 1}{\sqrt{\frac{n}{K}}\Delta_{k_0+1} - 1}\right)$$
$$\leqslant \frac{16K}{n\Delta_{k_0+1}} \log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \left(\frac{75}{74} + \frac{\sqrt{75}}{\sqrt{75}-1}\right)\frac{K}{n\Delta_{k_0+1}}$$

where the penultimate inequality uses $\Delta_{k_0+1} \geqslant \sqrt{\frac{75K}{n}}$ and $\log(1+x) \leqslant x$ for any $x \geqslant 0$.

Gathering the results of the three steps, we get

$$\begin{aligned}
\overline{R}_n &\leqslant n\Delta_{k_0} + \sum_{k=k_0+1}^{K} \left(1 + 7\frac{\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k} + \frac{\exp(-14c^2\log(75))}{2c^2(1-c^2)\Delta_k}\right) \\
&\qquad + \frac{16K}{\Delta_{k_0+1}} \log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \left(\frac{75}{74} + \frac{\sqrt{75}}{\sqrt{75}-1}\right)\frac{K}{\Delta_{k_0+1}} \\
&\leqslant n\Delta_{k_0} + K + (16+7)K\frac{\log\left(\frac{n}{K}\Delta_{k_0+1}^2\right)}{\Delta_{k_0+1}} + (16+16)\frac{K}{\Delta_{k_0+1}} \\
&\leqslant n\delta_0 \mathbb{1}_{\Delta \leqslant \delta_0} + 23K\frac{\log\left(\frac{n}{K}\Delta_{k_0+1}^2\right)}{\Delta_{k_0+1}} + \frac{33K}{\Delta_{k_0+1}} \\
&\leqslant 23K\frac{\log\left(\frac{n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)} + \frac{108K}{\max(\Delta,\delta_0)} \\
&\leqslant 23K\frac{\log\left(\frac{110n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)},
\end{aligned}$$

which implies (3.13) and also $\overline{R}_n \leqslant 24\sqrt{nK}$. Since Proposition 2.1 implies $\mathbb{E}R_n - \overline{R}_n \leqslant \sqrt{nK}$, we have proved (3.14). For (3.15), Proposition 2.1 also implies

$$\mathbb{E}R_n - \overline{R}_n \leqslant \min\left(\frac{K}{2\Delta}, \frac{\sqrt{nK}}{2}\right) \leqslant \frac{K\sqrt{75}}{2\max(\Delta,\delta_0)},$$

which implies

$$\mathbb{E}R_n \leqslant 23K\frac{\log\left(\frac{133n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)}.$$

CHAPTER 4

$\mathcal{X}$-Armed Bandits

We consider a generalization of stochastic bandits where the set of arms, $\mathcal{X}$, is allowed to be a generic measurable space and the mean-payoff function is "locally Lipschitz" with respect to a dissimilarity function that is known to the decision maker. Under this condition we construct an arm selection policy whose regret improves upon previous results for a large class of problems. In particular, our results imply that if $\mathcal{X}$ is the unit hypercube in a Euclidean space and the mean-payoff function has a finite number of global maxima around which the behavior of the function is locally Hölder with a known exponent, then the expected regret is bounded up to a logarithmic factor by $\sqrt{n}$, i.e., the rate of the growth of the regret is independent of the dimension of the space. We also prove the minimax optimality of our algorithm when the dissimilarity is a metric.

Contents

This chapter is a joint work with Rémi Munos, Gilles Stoltz and Csaba Szepesvari. It is based on the extended version Bubeck et al. [2009d] (currently under submission) of Bubeck et al. [2009c] which appeared in Advances in Neural Information Processing Systems 22.

1. Introduction

In the classical stochastic bandit problem, described in Chapter 2, a gambler tries to maximize his revenue by sequentially playing one of a finite number of slot machines that are associated with initially unknown (and potentially different) payoff distributions [Robbins, 1952]. Assuming

old-fashioned slot machines, the gambler pulls the arms of the machines one by one in a sequential manner, simultaneously learning about the machines' payoff-distributions and gaining actual monetary reward. Thus, in order to maximize his gain, the gambler must choose the next arm by taking into consideration both the urgency of gaining reward ("exploitation") and acquiring new information ("exploration").

Maximizing the total cumulative payoff is equivalent to minimizing the (total) *regret*, i.e., minimizing the difference between the total cumulative payoff of the gambler and the one of another clairvoyant gambler who chooses the arm with the best mean-payoff in every round. The quality of the gambler's strategy can be characterized as the rate of growth of his expected regret with time. In particular, if this rate of growth is sublinear, the gambler in the long run plays as well as the clairvoyant gambler. In this case the gambler's strategy is called Hannan consistent.

Bandit problems have been studied in the Bayesian framework [Gittins, 1989], as well as in the frequentist parametric [Lai and Robbins, 1985, Agrawal, 1995a] and non-parametric settings [Auer et al., 2002], and even in non-stochastic scenarios Auer et al. [2003], Cesa-Bianchi and Lugosi [2006]. While in the Bayesian-case the question is how to play optimally (i.e., the problem is really a computational problem), in the frequentist case the question is how to achieve low rate of growth of the regret in the lack of prior information, i.e., it is a statistical question. In this chapter we consider the stochastic, frequentist, non-parametric setting.

Although the first papers studied bandits with a finite number of arms, researchers have soon realized that bandits with infinitely many arms are also interesting, as well as practically significant. One particularly important case is when the arms are identified by a finite number of continuous-valued parameters, resulting in *online optimization* problems over continuous finite-dimensional spaces. Such problems are ubiquitous to operations research and control. Examples are "pricing a new product with uncertain demand in order to maximize revenue, controlling the transmission power of a wireless communication system in a noisy channel to maximize the number of bits transmitted per unit of power, and calibrating the temperature or levels of other inputs to a reaction so as to maximize the yield of a chemical process" [Cope, 2004]. Other examples are optimizing parameters of schedules, rotational systems, traffic networks or online parameter tuning of numerical methods. During the last decades numerous authors have investigated such "continuum-armed" bandit problems [Agrawal, 1995b, Kleinberg, 2004, Auer et al., 2007, Kleinberg et al., 2008a, Cope, 2004]. A special case of interest, which forms a bridge between the case of a finite number of arms and the continuum-armed setting, is formed by bandit linear optimization, see [Dani et al., 2008] and the references therein.

In many of the above-mentioned problems, however, the natural domain of some of the optimization parameters is a discrete set, while other parameters are still continuous-valued. For example, in the pricing problem different product lines could also be tested while tuning the price, or in the case of transmission power control different protocols could be tested while optimizing the power. In other problems, such as in online sequential search, the parameter-vector to be optimized is an infinite sequence over a finite alphabet [Coquelin and Munos, 2007].

The motivation for this chapter is to handle all these various cases in a unified framework. More precisely, we consider a general setting that allows us to study bandits with almost no restriction on the set of arms. In particular, we allow the set of arms to be an arbitrary measurable space. Since we allow non-denumerable sets, we shall assume that the gambler has some knowledge about the behavior of the mean-payoff function (in terms of its local regularity around its maxima, roughly speaking). This is because when the set of arms is uncountably infinite and absolutely no assumptions are made on the payoff function, it is impossible to construct a strategy that simultaneously achieves sublinear regret for all bandits problems. When the set of arms is a

continuous metric space previous works have assumed either the global smoothness of the payoff function [Agrawal, 1995b, Kleinberg, 2004, Kleinberg et al., 2008a, Cope, 2004] or local smoothness in the vicinity of the maxima Auer et al. [2007]. These smoothness assumptions are indeed reasonable in many practical problems of interest.

In this chapter, we assume that there exists a dissimilarity function that constraints the behavior of the mean-payoff function. In particular, the dissimilarity function is assumed to locally set a bound on the decrease of the mean-payoff function at each of its global maxima. We also assume that the decision maker can construct a recursive covering of the space of arms in such a way that the diameters of the sets in the covering shrink at a known geometric rate when measured with this dissimilarity.

Our work generalizes and improves previous works on continuum-armed bandits. Kleinberg [2004] and Auer et al. [2007] focused on one-dimensional problems. Recently, Kleinberg et al. [2008a] considered generic metric spaces assuming that the mean-payoff function is Lipschitz with respect to the (known) metric of the space. They proposed a novel algorithm that achieves essentially the best possible regret in a minimax sense with respect to these environments. The goal of this chapter is to further these works in a number of ways:

(i): we allow the set of arms to be a generic measurable space;
(ii): we propose a practical algorithm motivated by the recent successful tree-based optimization algorithms Kocsis and Szepesvari [2006], Gelly et al. [2006], Coquelin and Munos [2007];
(iii): we show that the algorithm is able to exploit higher order of smoothness.

In particular, as we shall argue in Section 5, (i) improves upon the results of Auer et al. [2007], while (i), (ii) and (iii) improve upon the work of Kleinberg et al. [2008a]. Compared to Kleinberg et al. [2008a], our work represents an improvement in the fact that just like Auer et al. [2007] we make use of the *local* properties of the mean-payoff function around the maxima only and do not assume a global property, such as Lipschitzness in the whole space. This allows us to obtain a regret which scales as $\widetilde{O}(\sqrt{n})$ [1] when, e.g., the space is the unit hypercube and the mean-payoff function is locally Hölder continuous with known exponent in the neighborhood of any global maximum (these global maxima being finite in number) and bounded away from the maxima outside of these neighborhoods. Thus, we get the desirable property that the rate of growth of the regret is independent of the dimensionality of the input space. We also prove a minimax lower bound that matches our upper bound up to logarithmic factors, showing that the performance of our algorithm is essentially unimprovable in a minimax sense. Besides these theoretical advances, the algorithm is anytime and easy to implement. Since it is based on ideas that have proved to be efficient in search and planning [Gelly and Silver, 2007, 2008, Schadd et al., 2008, Chaslot et al., 2008, Finnsson and Bjornsson, 2008], we expect it to perform equally well in practice and to make a significant impact on how online global optimization is performed.

Outline.

(1) In Section 2 we formalize the $\mathcal{X}$–armed bandit problem.
(2) In Section 3 we describe the basic strategy proposed, called HOO (*hierarchical optimistic optimization*).
(3) We present the main results in Section 4. We start by specifying and explaining our assumptions (Section 4.1) under which various regret bounds are proved. Then we prove

[1] We write $u_n = \widetilde{O}(v_n)$ when $u_n = O(v_n)$ up to a logarithmic factor.

a distribution-dependent bound for the basic version of HOO (Section 4.2). A problem with the basic algorithm is that its computational cost increases quadratically with the number of time steps. Assuming the knowledge of the horizon, we thus propose a computationally more efficient variant of the basic algorithm, called *truncated HOO* and prove that it enjoys a regret bound identical to the one of the basic version (Section 4.3) while its computational complexity is only log-linear in the number of time steps. The first set of assumptions constrains the mean-payoff function everywhere. A second set of assumptions is therefore presented that puts constraints on the mean-payoff function only in a small vicinity of its global maxima; we then propose another algorithm, called *local HOO*, which is proven to enjoy a regret again essentially similar to the one of the basic version (Section 4.4). Finally, we prove the minimax optimality of HOO in metric spaces (Section 4.5).

(4) In Section 5 we compare the results of this chaper with previous works.

2. Problem setup

A *stochastic bandit problem* $\mathcal{B}$ is a pair $\mathcal{B} = (\mathcal{X}, M)$, where $\mathcal{X}$ is a measurable space of arms and M determines the distribution of rewards associated with each arm. We say that M is a *bandit environment* on $\mathcal{X}$. Formally, M is an mapping $\mathcal{X} \to \mathcal{M}_1(\mathbb{R})$, where $\mathcal{M}_1(\mathbb{R})$ is the space of probability distributions over the reals. The distribution assigned to arm $x \in \mathcal{X}$ is denoted by M_x. We require that for each arm $x \in \mathcal{X}$, the distribution M_x admits a first-order moment; we then denote by $f(x)$ its expectation ("mean payoff"),

$$f(x) = \int y \; dM_x(y) \, .$$

The mean-payoff function f thus defined is assumed to be measurable. For simplicity, we shall also assume that all M_x have bounded supports, included in some fixed bounded interval[2], say, the unit interval $[0, 1]$. Then, f also takes bounded values, in $[0, 1]$.

A decision maker (the gambler of the introduction) that interacts with a stochastic bandit problem $\mathcal{B}$ plays a game at discrete time steps according to the following rules. In the first round the decision maker can select an arm $X_1 \in \mathcal{X}$ and receives a reward Y_1 drawn at random from M_{X_1}. In round $n > 1$ the decision maker can select an arm $X_n \in \mathcal{X}$ based on the information available up to time n, i.e., $(X_1, Y_1, \ldots, X_{n-1}, Y_{n-1})$, and receives a reward Y_n drawn from M_{X_n}, independently of $(X_1, Y_1, \ldots, X_{n-1}, Y_{n-1})$ given X_n. Note that a decision maker may randomize his choice, but can only use information available up to the point in time when the choice is made.

Formally, a *strategy of the decision maker* in this game ("bandit strategy") can be described by an infinite sequence of measurable mappings, $\varphi = (\varphi_1, \varphi_2, \ldots)$, where φ_n maps the space of past observations,

$$\mathcal{H}_n = \big(\mathcal{X} \times [0, 1]\big)^{n-1},$$

to the space of probability measures over $\mathcal{X}$. By convention, φ_1 does not take any argument. A strategy is called *deterministic* if for every n, φ_n is a Dirac distribution.

The goal of the decision maker is to maximize his expected cumulative reward. Equivalently, the goal can be expressed as minimizing the expected cumulative regret, which is defined as follows. Let

$$f^* = \sup_{x \in \mathcal{X}} f(x)$$

[2]More generally, our results would also hold when the tails of the reward distributions are uniformly sub-Gaussian.

be the best expected payoff in a single round. At round n, the *cumulative regret* of a decision maker playing $\mathcal{B}$ is

$$R_n = n\, f^* - \sum_{t=1}^{n} f(X_t)\,.$$

Note that this definition differs from the ones in Chapter 2 and 3. Indeed, in stochastic bandits games the distinction between regret and pseudo-regret is rather unnatural. From now we shall call $\mathbb{E}[R_n]$ the expected regret (rather than the pseudo-regret) and focus on this quantity.

REMARK 4.1. *As it is argued in Chapter 6, in many real-world problems, the decision maker is not interested in his cumulative regret but rather in its simple regret. The latter can be defined as follows. After n rounds of play in a stochastic bandit problem $\mathcal{B}$, the decision maker is asked to make a recommendation $Z_n \in \mathcal{X}$ based on the n obtained rewards $Y_1, \dots, Y_n$. The simple regret of this recommendation equals*

$$r_n = f^* - f(Z_n)\,.$$

In this chaper we focus on the cumulative regret R_n, but all the results can be readily extended to the simple regret by considering the recommendation $Z_n = X_{T_n}$, where T_n is drawn uniformly at random in $\{1, \dots, n\}$. Indeed, in this case,

$$\mathbb{E}\big[r_n\big] \leqslant \frac{\mathbb{E}\big[R_n\big]}{n}.$$

3. The Hierarchical Optimistic Optimization (HOO) strategy

The HOO strategy (cf. Algorithm 1) incrementally builds an estimate of the mean-payoff function f over $\mathcal{X}$. The core idea (as in previous works) is to estimate f precisely around its maxima, while estimating it loosely in other parts of the space $\mathcal{X}$. To implement this idea, HOO maintains a binary tree whose nodes are associated with measurable regions of the arm-space $\mathcal{X}$ such that the regions associated with nodes deeper in the tree (further away from the root) represent increasingly smaller subsets of $\mathcal{X}$. The tree is built in an incremental manner. At each node of the tree, HOO stores some statistics based on the information received in previous rounds. In particular, HOO keeps track of the number of times a node was traversed up to round n and the corresponding empirical average of the rewards received so far. Based on these, HOO assigns an optimistic estimate (denoted by B) to the maximum mean-payoff associated with each node. These estimates are then used to select the next node to "play". This is done by traversing the tree, beginning from the root, and always following the node with the highest B–value (cf. lines 4–14 of Algorithm 1). Once a node is selected, a point in the region associated with it is chosen (line 16) and is sent to the environment. Based on the point selected and the received reward, the tree is updated (lines 18–33).

The tree of coverings which HOO needs to receive as an input is an infinite binary tree whose nodes are associated with subsets of $\mathcal{X}$. The nodes in this tree are indexed by pairs of integers (h, i); node (h, i) is located at depth $h \geqslant 0$ from the root. The range of the second index, i, associated with nodes at depth h is restricted by $1 \leqslant i \leqslant 2^h$. Thus, the root node is denoted by $(0, 1)$. By convention, $(h+1, 2i-1)$ and $(h+1, 2i)$ are used to refer to the two children of the node (h, i). Let $\mathcal{P}_{h,i} \subset \mathcal{X}$ be the region associated with node (h, i). By assumption, these regions are measurable and must satisfy the constraints

(4.1a) $\qquad \mathcal{P}_{0,1} = \mathcal{X}\,,$

(4.1b) $\qquad \mathcal{P}_{h,i} = \mathcal{P}_{h+1,2i-1} \cup \mathcal{P}_{h+1,2i}\,, \qquad$ for all $h \geqslant 0$ and $1 \leqslant i \leqslant 2^h$.

As a corollary, the regions $\mathcal{P}_{h,i}$ at any level $h \geqslant 0$ cover the space $\mathcal{X}$,

$$\mathcal{X} = \bigcup_{i=1}^{2^h} \mathcal{P}_{h,i} \,,$$

explaining the term "tree of coverings".

In the algorithm listing the recursive computation of the B–values (lines 28–33) makes a local copy of the tree; of course, this part of the algorithm could be implemented in various other ways. Other arbitrary choices in the algorithm as shown here are how tie breaking in the node selection part is done (lines 9–12), or how a point in the region associated with the selected node is chosen (line 16). We note in passing that implementing these differently would not change our results.

To facilitate the formal study of the algorithm, we shall need some more notation. In particular, we shall introduce time-indexed versions ($\mathcal{T}_n$, (H_n, I_n), X_n, Y_n, $\widehat{\mu}_{h,i}(n)$, etc.) of the quantities used by the algorithm. The convention used is that the indexation by n is used to indicate the value taken at the end of the n–th round.

In particular, $\mathcal{T}_n$ is used to denote the finite subtree stored by the algorithm at the end of round n. Thus, the initial tree is $\mathcal{T}_0 = \{(0,1)\}$ and it is expanded round after round as

$$\mathcal{T}_n = \mathcal{T}_{n-1} \cup \{(H_n, I_n)\} \,,$$

where (H_n, I_n) is the node selected in line 16. We call (H_n, I_n) *the node played in round* n. We use X_n to denote the point selected by HOO in the region associated with the node played in round n, while Y_n denotes the received reward.

Node selection works by comparing B–values and always choosing the node with the highest B–value. The B–value, $B_{h,i}(n)$, at node (h,i) by the end of round n is an estimated upper bound on the mean-payoff function at node (h,i). To define it we first need to introduce the average of the rewards received in rounds when some descendant of node (h,i) was chosen (by convention, each node is a descendant of itself):

$$\widehat{\mu}_{h,i}(n) = \frac{1}{T_{h,i}(n)} \sum_{t=1}^{n} Y_t \, \mathbb{I}_{\{(H_t,I_t) \in \mathcal{C}(h,i)\}} \,.$$

Here, $\mathcal{C}(h,i)$ denotes the set of all descendants of a node (h,i) in the infinite tree,

$$\mathcal{C}(h,i) = \big\{(h,i)\big\} \cup \mathcal{C}(h+1, 2i-1) \cup \mathcal{C}(h+1, 2i) \,,$$

and $T_{h,i}(n)$ is the number of times a descendant of (h,i) is played up to and including round n, that is,

$$T_{h,i}(n) = \sum_{t=1}^{n} \mathbb{I}_{\{(H_t,I_t) \in \mathcal{C}(h,i)\}} \,.$$

A key quantity determining $B_{h,i}(n)$ is $U_{h,i}(n)$, an initial estimate of the maximum of the mean-payoff function in the region $\mathcal{P}_{h,i}$ associated with node (h,i):

$$U_{h,i}(n) = \begin{cases} \widehat{\mu}_{h,i}(n) + \sqrt{\dfrac{2 \ln n}{T_{h,i}(n)}} + \nu_1 \rho^h, & \text{if } T_{h,i}(n) > 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.2}$$

In the expression corresponding to the case $T_{h,i}(n) > 0$, the first term added to the average of rewards accounts for the uncertainty arising from the randomness of the rewards that the average is based on, while the second term, $\nu_1 \rho^h$, accounts for the maximum possible variation of the mean-payoff function over the region $\mathcal{P}_{h,i}$. The actual bound on the maxima used in HOO is

Algorithm 1 The HOO strategy

Parameters: Two real numbers $\nu_1 > 0$ and $\rho \in (0,1)$, a sequence $(\mathcal{P}_{h,i})_{h\geqslant 0, 1\leqslant i\leqslant 2^h}$ of subsets of $\mathcal{X}$ satisfying the conditions (4.1a) and (4.1b).
Auxiliary function LEAF($\mathcal{T}$): outputs a leaf of $\mathcal{T}$.
Initialization: $\mathcal{T} = \{(0,1)\}$ and $B_{1,2} = B_{2,2} = +\infty$.

for $n = 1, 2, \ldots$ **do** ▷ Strategy HOO in round $n \geqslant 1$
  $(h,i) \leftarrow (0,1)$ ▷ Start at the root
  $P \leftarrow \{(h,i)\}$ ▷ P stores the path traversed in the tree
  while $(h,i) \in \mathcal{T}$ **do** ▷ Search the tree $\mathcal{T}$
    if $B_{h+1,2i-1} > B_{h+1,2i}$ **then** ▷ Select the "more promising" child
      $(h,i) \leftarrow (h+1, 2i-1)$
    else if $B_{h+1,2i-1} < B_{h+1,2i}$ **then**
      $(h,i) \leftarrow (h+1, 2i)$
    else ▷ Tie-breaking rule
      $Z \sim \mathrm{Ber}(0.5)$ ▷ e.g., choose a child at random
      $(h,i) \leftarrow (h+1, 2i-Z)$
    end if
    $P \leftarrow P \cup \{(h,i)\}$
  end while
  $(H,I) \leftarrow (h,i)$ ▷ The selected node
  Choose arm X in $\mathcal{P}_{H,I}$ and play it ▷ Arbitrary selection of an arm
  Receive corresponding reward Y
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{(H,I)\}$ ▷ Extend the tree
  for all $(h,i) \in P$ **do** ▷ Update the statistics T and $\hat{\mu}$ stored in the path
    $T_{h,i} \leftarrow T_{h,i} + 1$ ▷ Increment the counter of node (h,i)
    $\hat{\mu}_{h,i} \leftarrow \big(1 - 1/T_{h,i}\big)\hat{\mu}_{h,i} + Y/T_{h,i}$ ▷ Update the mean $\hat{\mu}_{h,i}$ of node (h,i)
  end for
  for all $(h,i) \in \mathcal{T}$ **do** ▷ Update the statistics U stored in the tree
    $U_{h,i} \leftarrow \hat{\mu}_{h,i} + \sqrt{(2\ln n)/T_{h,i}} + \nu_1 \rho^h$ ▷ Update the U–value of node (h,i)
  end for
  $B_{H+1,2I-1} \leftarrow +\infty$ ▷ B–values of the children of the new leaf
  $B_{H+1,2I} \leftarrow +\infty$
  $\mathcal{T}' \leftarrow \mathcal{T}$ ▷ Local copy of the current tree $\mathcal{T}$
  while $\mathcal{T}' \neq \{(0,1)\}$ **do** ▷ Backward computation of the B–values
    $(h,i) \leftarrow \mathrm{LEAF}(\mathcal{T}')$ ▷ Take any remaining leaf
    $B_{h,i} \leftarrow \min\Big\{U_{h,i},\, \max\big\{B_{h+1,2i-1}, B_{h+1,2i}\big\}\Big\}$ ▷ Backward computation
    $\mathcal{T}' \leftarrow \mathcal{T}' \setminus \{(h,i)\}$ ▷ Drop updated leaf (h,i)
  end while
end for

defined recursively by

$$B_{h,i}(n) = \begin{cases} \min\Big\{U_{h,i}(n),\, \max\big\{B_{h+1,2i-1}(n), B_{h+1,2i}(n)\big\}\Big\}, & \text{if } (h,i) \in \mathcal{T}_n; \\ +\infty, & \text{otherwise.} \end{cases}$$

The role of $B_{h,i}(n)$ is to put a tight, optimistic, high-probability upper bound on the best mean-payoff that can be achieved in the region $\mathcal{P}_{h,i}$. By assumption, $\mathcal{P}_{h,i} = \mathcal{P}_{h+1,2i-1} \cup \mathcal{P}_{h+1,2i}$. Thus, assuming that $B_{h+1,2i-1}(n)$ (resp., $B_{h+1,2i}(n)$) is a valid upper bound for region $\mathcal{P}_{h+1,2i-1}$ (resp., $\mathcal{P}_{h+1,2i}$), we see that $\max\big\{B_{h+1,2i-1}(n), B_{h+1,2i}(n)\big\}$ must be a valid upper bound for

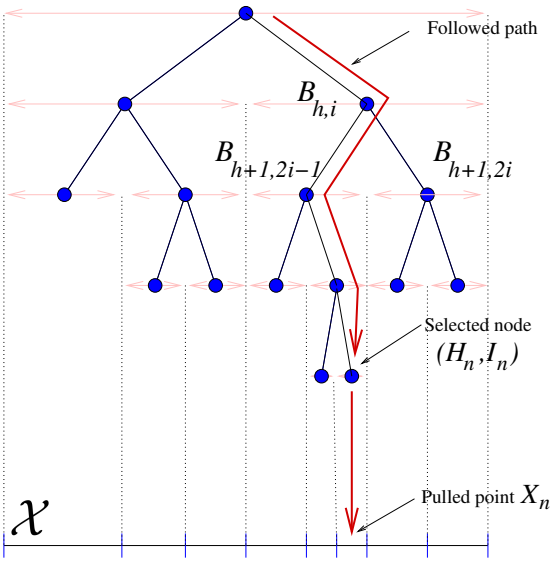

Figure 1: Illustration of the node selection procedure in round n. The tree represents $\mathcal{T}_n$. In the illustration, $B_{h+1,2i-1}(n-1) > B_{h+1,2i}(n-1)$, therefore, the selected path included the node $(h+1, 2i-1)$ rather than the node $(h+1, 2i)$.

region $\mathcal{P}_{h,i}$. Since $U_{h,i}(n)$ is another valid upper bound for region $\mathcal{P}_{h,i}$, we get a tighter upper bound by taking the minimum of these bounds.

Obviously, for leafs (h, i) of the tree $\mathcal{T}_n$, one has $B_{h,i}(n) = U_{h,i}(n)$, while close to the root one may expect that $B_{h,i}(n) < U_{h,i}(n)$; that is, the upper bounds close to the root are expected to be less biased than the ones associated with nodes farther away from the root.

Note that at the beginning of round n, the algorithm uses $B_{h,i}(n-1)$ to select the node (H_n, I_n) to be played (since $B_{h,i}(n)$ will only be available at the end of round n). It does so by following a path from the root node to an inner node with only one child or a leaf and finally considering a child (H_n, I_n) of the latter; at each node of the path, the child with highest B–value is chosen, till the node (H_n, I_n) with infinite B–value is reached.

Illustrations. Figure 1 illustrates the computation done by HOO in round n, as well as the correspondence between the nodes of the tree constructed by the algorithm and their associated regions. Figure 2 shows trees built by running HOO for a specific environment.

Computational complexity. At the end of round n, the size of the active tree $\mathcal{T}_n$ is at most n, making the storage requirements of HOO linear in n. In addition, the statistics and B–values of all nodes in the active tree need to be updated, which thus takes time $O(n)$. HOO runs in time $O(n)$ at each round n, making the algorithm's total running time up to round n quadratic in n. In Section 4.3 we modify HOO so that if the time horizon n_0 is known in advance, the total running time is $O(n_0 \ln n_0)$, while the modified algorithm will be shown to enjoy essentially the same regret bound as the original version.

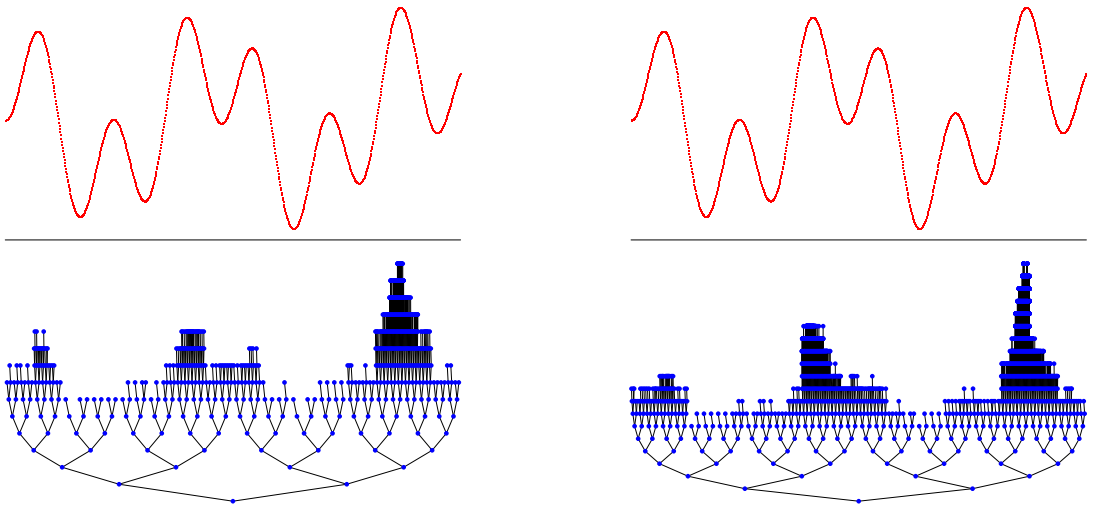

Figure 2: The trees (bottom figures) built by HOO after 1,000 (left) and 10,000 (right) rounds. The mean-payoff function (shown in the top part of the figure) is $x \in [0,1] \longmapsto 1/2\big(\sin(13x)\sin(27x)+1\big)$; the corresponding payoffs are Bernoulli-distributed. The inputs of HOO are as follows: the tree of coverings is formed by all dyadic intervals, $\nu_1 = 1$ and $\rho = 1/2$. The tie-breaking rule is to choose a child at random (as shown in the Algorithm 1), while the points in $\mathcal{X}$ to be played are chosen as the centers of the dyadic intervals. Note that the tree is extensively refined where the mean-payoff function is near-optimal, while it is much less developed in other regions.

4. Main results

We start by describing and commenting the assumptions that we need to analyze the regret of HOO. This is followed by stating the first upper bound, followed by some improvements on the basic algorithm. The section is finished by the statement of our results on the minimax optimality of HOO.

4.1. Assumptions. The main assumption will concern the "smoothness" of the mean-payoff function. However, somewhat unconventionally, we shall use a notion of smoothness that is built around dissimilarity functions rather than distances, allowing us to deal with function classes of highly different smoothness orders in a unified manner. Before stating our smoothness assumptions, we define the notion of a dissimilarity function and some associated concepts.

DEFINITION 4.1 (Dissimilarity). *A* dissimilarity ℓ *over* $\mathcal{X}$ *is a non-negative mapping* $\ell : \mathcal{X}^2 \to \mathbb{R}$ *satisfying* $\ell(x,x) = 0$ *for all* $x \in \mathcal{X}$.

Given a dissimilarity ℓ, the *diameter* of a subset A of $\mathcal{X}$ as measured by ℓ is defined by

$$\operatorname{diam}(A) = \sup_{x,y \in A} \ell(x,y)\,,$$

while the ℓ*–open ball* of $\mathcal{X}$ with radius $\varepsilon > 0$ and center $x \in \mathcal{X}$ is defined by

$$\mathcal{B}(x,\varepsilon) = \{\, y \in \mathcal{X} \,:\, \ell(x,y) < \varepsilon \,\}\,.$$

Note that the dissimilarity ℓ will be mostly used in the theoretical analysis of HOO; however, by carefully choosing the parameters of HOO (the tree of coverings and the real numbers $\nu_1 > 0$ and $\rho < 1$) for the (set of) two assumptions below to be satisfied, the user of the algorithm effectively chooses a dissimilarity. He however does not have to construct it explicitly.

Assumptions Given the parameters of HOO, that is, the real numbers $\nu_1 > 0$ and $\rho \in (0,1)$ and the tree of coverings $(\mathcal{P}_{h,i})$, there exists a dissimilarity function ℓ such that the following two assumptions are satisfied.

A1. There exists $\nu_2 > 0$ such that for all integers $h \geqslant 0$,
  (a) $\operatorname{diam}(\mathcal{P}_{h,i}) \leqslant \nu_1 \rho^h$ for all $i = 1, \dots, 2^h$;
  (b) for all $i = 1, \dots, 2^h$, there exists $x^\circ_{h,i} \in \mathcal{P}_{h,i}$ such that

$$\mathcal{B}_{h,i} \stackrel{\text{def}}{=} \mathcal{B}\big(x^\circ_{h,i},\, \nu_2 \rho^h\big) \subset \mathcal{P}_{h,i}\,;$$

  (c) $\mathcal{B}_{h,i} \cap \mathcal{B}_{h,j} = \emptyset$ for all $1 \leqslant i < j \leqslant 2^h$.

A2. The mean-payoff function f satisfies that for all $x, y \in \mathcal{X}$,

$$f^* - f(y) \leqslant f^* - f(x) + \max\big\{f^* - f(x),\, \ell(x,y)\big\}\,. \tag{4.3}$$

Assumption A1 ensures that the regions in the tree of coverings $(\mathcal{P}_{h,i})$ shrink exactly at a geometric rate. The following example shows how to satisfy A1 when the domain $\mathcal{X}$ is a D–dimensional hyper-rectangle and the dissimilarity is some positive power of the Euclidean (or supremum) norm.

EXAMPLE 4.1. *Assume that $\mathcal{X}$ is a D-dimension hyper-rectangle and consider the dissimilarity $\ell(x,y) = b\|x-y\|_2^a$, where $a > 0$ and $b > 0$ are real numbers and $\|\cdot\|_2$ is the Euclidean norm. Define the tree of coverings $(\mathcal{P}_{h,i})$ in the following inductive way: let $\mathcal{P}_{0,1} = \mathcal{X}$. Given a node $\mathcal{P}_{h,i}$, let $\mathcal{P}_{h+1,2i-1}$ and $\mathcal{P}_{h+1,2i}$ be obtained from the hyper-rectangle $\mathcal{P}_{h,i}$ by splitting it in the middle along its longest side (ties can be broken arbitrarily).*

We now argue that Assumption A1 is satisfied. With no loss of generality we take $\mathcal{X} = [0,1]^D$. Then, for all integers $u \geqslant 0$ and $0 \leqslant k \leqslant D-1$,

$$\operatorname{diam}(\mathcal{P}_{uD+k,1}) = b\left(\frac{1}{2^u}\sqrt{D - \frac{3}{4}\,k}\right)^a \leqslant b\left(\frac{\sqrt{D}}{2^u}\right)^a\,.$$

It is now easy to see that Assumption A1 is satisfied for the indicated dissimilarity, e.g., with the choice of the parameters $\rho = 2^{-a/D}$ and $\nu_1 = b\left(2\sqrt{D}\right)^a$ for HOO, and the value $\nu_2 = b/2^a$.

EXAMPLE 4.2. *In the same setting, with the same tree of coverings $(\mathcal{P}_{h,i})$ over $\mathcal{X} = [0,1]^D$, but now with the dissimilarity $\ell(x,y) = b\|x-y\|_\infty^a$, we get that for all integers $u \geqslant 0$ and $0 \leqslant k \leqslant D-1$,*

$$\operatorname{diam}(\mathcal{P}_{uD+k,1}) = b\left(\frac{1}{2^u}\right)^a\,.$$

This time, Assumption A1 is satisfied, e.g., with the choice of the parameters $\rho = 2^{-a/D}$ and $\nu_1 = b\,2^a$ for HOO, and the value $\nu_2 = b/2^a$.

The second assumption, A2, concerns the environment; when Assumption A2 is satisfied, we say that f is *weakly Lipschitz* with respect to (w.r.t.) ℓ. The choice of this terminology follows from the fact that if f is 1–Lipschitz w.r.t. ℓ, i.e., for all $x, y \in \mathcal{X}$, one has $|f(x) - f(y)| \leqslant \ell(x,y)$, then it is also weakly Lipschitz w.r.t. ℓ.

On the other hand, weak Lipschitzness is a milder requirement. It implies local (one-sided) 1–Lipschitzness at any global maximum, since at any arm x^* such that $f(x^*) = f^*$, the criterion (4.3) rewrites to $f(x^*) - f(y) \leqslant \ell(x^*, y)$. In the vicinity of other arms x, the constraint is milder as the arm x gets worse (as $f^* - f(x)$ increases) since the condition (4.3) rewrites to

$$\forall\, y \in \mathcal{X}, \qquad f(x) - f(y) \leqslant \max\big\{f^* - f(x),\, \ell(x,y)\big\}\,.$$

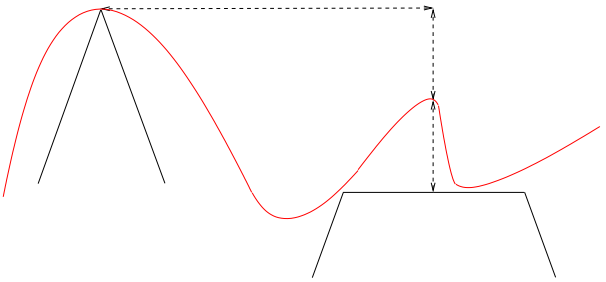

Figure 3: Illustration of the property of weak Lipschitzness.

Here is another interpretation of these two facts; it will be useful when considering local assumptions in Section 4.4 (a weaker set of assumptions). First, concerning the behavior around global maxima, Assumption A2 implies that for any set $\mathcal{A} \subset \mathcal{X}$ with $\sup_{x\in\mathcal{A}} f(x) = f^*$,

$$f^* - \inf_{x\in\mathcal{A}} f(x) \leqslant \operatorname{diam}(\mathcal{A}). \tag{4.4}$$

Second, it can be seen that Assumption A2 is equivalent[3] to the following property: for all $x \in \mathcal{X}$ and $\varepsilon \geqslant 0$,

$$\mathcal{B}\big(x,\, f^* - f(x) + \varepsilon\big) \;\subset\; \mathcal{X}_{2\left(f^* - f(x)\right)+\varepsilon} \tag{4.5}$$

where

$$\mathcal{X}_\varepsilon = \big\{x \in \mathcal{X} : f(x) \geqslant f^* - \varepsilon\big\}$$

denotes the set of ε*–optimal arms*. This second property essentially states that there is no sudden and large drop in the mean-payoff function (note that this property can be satisfied even for discontinuous functions).

Figure 3 presents an illustration of the two properties discussed above.

Before stating our main results, we provide a straightforward, though useful consequence of Assumptions A1 and A2, which should be seen as an intuitive justification for the third term in (4.2).

For all nodes (h, i), let

$$f^*_{h,i} = \sup_{x\in\mathcal{P}_{h,i}} f(x) \qquad \text{and} \qquad \Delta_{h,i} = f^* - f^*_{h,i}\,.$$

$\Delta_{h,i}$ is called the *suboptimality factor* of node (h, i). Depending whether it is positive or not, a node (h, i) is called *suboptimal* ($\Delta_{h,i} > 0$) or *optimal* ($\Delta_{h,i} = 0$).

LEMMA 4.1. *Under Assumptions A1 and A2, if the suboptimality factor $\Delta_{h,i}$ of a region $\mathcal{P}_{h,i}$ is bounded by $c\nu_1\rho^h$ for some $c \geqslant 0$, then all arms in $\mathcal{P}_{h,i}$ are $\max\{2c, c+1\}\,\nu_1\rho^h$–optimal, that is,*

$$\mathcal{P}_{h,i} \subset \mathcal{X}_{\max\{2c,c+1\}\,\nu_1\rho^h}\,.$$

PROOF. For all $\delta > 0$, we denote by $x^*_{h,i}(\delta)$ an element of $\mathcal{P}_{h,i}$ such that

$$f\big(x^*_{h,i}(\delta)\big) \geqslant f^*_{h,i} - \delta = f^* - \Delta_{h,i} - \delta\,.$$

[3]That Assumption A2 implies (4.5) is immediate; for the converse, it suffices to consider, for each $y \in \mathcal{X}$, the sequence

$$\varepsilon_n = \Big(\ell(x, y) - \big(f^* - f(x)\big)\Big)_+ + 1/n\,,$$

where $(\,\cdot\,)_+$ denotes the nonnegative part.

By the weak Lipschitz property (Assumption A2), it then follows that for all $y \in \mathcal{P}_{h,i}$,

$$f^* - f(y) \leqslant f^* - f\big(x^*_{h,i}(\delta)\big) + \max\Big\{f^* - f\big(x^*_{h,i}(\delta)\big),\, \ell\big(x^*_{h,i}(\delta),\, y\big)\Big\}$$
$$\leqslant \Delta_{h,i} + \delta + \max\big\{\Delta_{h,i} + \delta,\, \operatorname{diam}\mathcal{P}_{h,i}\big\}\,.$$

Letting $\delta \to 0$ and substituting the bounds on the suboptimality and on the diameter of $\mathcal{P}_{h,i}$ (Assumption A1) concludes the proof. □

4.2. Upper bound for the regret of HOO. Auer et al. [Auer et al., 2007, Assumption 2] observed that the regret of a continuum-armed bandit algorithm should depend on how fast the volumes of the sets of ε–optimal arms shrink as $\varepsilon \to 0$. Here, we capture this by defining a new notion, the near-optimality dimension of the mean-payoff function. The connection between these concepts, as well as with the zooming dimension defined by Kleinberg et al. [2008a], will be further discussed in Section 5. We start by recalling the definition of packing numbers.

DEFINITION 4.2 (Packing number). *The ε–packing number $\mathcal{N}(\mathcal{X}, \ell, \varepsilon)$ of $\mathcal{X}$ w.r.t. the dissimilarity ℓ is the size of the largest packing of $\mathcal{X}$ with disjoint ℓ–open balls of radius ε. That is, $\mathcal{N}(\mathcal{X}, \ell, \varepsilon)$ is the largest integer k such that there exists k disjoint ℓ–open balls with radius ε contained in $\mathcal{X}$.*

We now define the c–near-optimality dimension, which characterizes the size of the sets $\mathcal{X}_{c\varepsilon}$ as a function of ε. It can be seen as some growth rate in ε of the metric entropy (measured in terms of ℓ and with packing numbers rather than covering numbers) of the set of $c\varepsilon$–optimal arms.

DEFINITION 4.3 (Near-optimality dimension). *For $c > 0$ the* c–near-optimality dimension *of f w.r.t. ℓ equals*

$$\max\left\{0,\, \limsup_{\varepsilon\to 0} \frac{\ln \mathcal{N}\big(\mathcal{X}_{c\varepsilon}, \ell, \varepsilon\big)}{\ln\big(\varepsilon^{-1}\big)}\right\}.$$

The following example shows that using a dissimilarity (rather than a metric, for instance) may sometimes allow for a significant reduction of the near-optimality dimension.

EXAMPLE 4.3. *Let $\mathcal{X} = [0,1]^D$ and let $f : [0,1]^D \to [0,1]$ be defined by $f(x) = 1 - \|x\|^a$ for some $a \geqslant 1$ and some norm $\|\cdot\|$ on $\mathbb{R}^D$. Consider the dissimilarity ℓ defined by $\ell(x,y) = \|x-y\|^a$. We shall see in Example 4.4 that f is weakly Lipschitz w.r.t. ℓ (in a sense however slightly weaker than the one given by* (4.4) *and* (4.5) *but sufficiently strong to ensure a result similar to the one of the main result, Theorem 4.1 below). Here we claim that the c–near-optimality dimension (for any $c > 0$) of f w.r.t. ℓ is 0. On the other hand, the c–near-optimality dimension (for any $c > 0$) of f w.r.t. the dissimilarity ℓ' defined, for $0 < b < a$, by $\ell'(x,y) = \|x-y\|^b$ is $(1/b - 1/a)D > 0$. In particular, when $a > 1$ and $b = 1$, the c–near-optimality dimension is $(1 - 1/a)D$.*

PROOF. **(sketch)** Fix $c > 0$. The set $\mathcal{X}_{c\varepsilon}$ is the $\|\cdot\|$–ball with center 0 and radius $(c\varepsilon)^{1/a}$, that is, the ℓ–ball with center 0 and radius $c\varepsilon$. Its ε–packing number w.r.t. ℓ is bounded by a constant depending only on D, c and a; hence, the value 0 for the near-optimality dimension w.r.t. the dissimilarity ℓ.

In case of ℓ', we are interested in the packing number of the $\|\cdot\|$–ball with center 0 and radius $(c\varepsilon)^{1/a}$ w.r.t. ℓ'–balls. The latter is of the order of

$$\left(\frac{(c\varepsilon)^{1/a}}{\varepsilon^{1/b}}\right)^D = c^{D/a}\big(\varepsilon^{-1}\big)^{(1/b-1/a)D}\,;$$

hence, the value $(1/b - 1/a)D$ for the near-optimality dimension in the case of the dissimilarity ℓ'.

Note that in all these cases the c–near-optimality dimension of f is independent of the value of c. □

We can now state our first main result. The proof is presented in Section 6.1.

THEOREM 4.1 (Regret bound for HOO). *Consider HOO tuned with parameters such that Assumptions A1 and A2 hold for some dissimilarity ℓ. Let d be the $4\nu_1/\nu_2$–near-optimality dimension of the mean-payoff function f w.r.t. ℓ. Then, for all $d' > d$, there exists a constant γ such that for all $n \geqslant 1$,*

$$\mathbb{E}\big[R_n\big] \leqslant \gamma\, n^{(d'+1)/(d'+2)} \big(\ln n\big)^{1/(d'+2)}.$$

Note that if d is infinite, then the bound is vacuous. The constant γ in the theorem depends on d' and on all other parameters of HOO and of the assumptions, as well as on the bandit environment M. The next section will exhibit a refined upper bound with a more explicit value of γ in terms of all these parameters.

REMARK 4.2. *The tuning of the parameters of HOO is critical for the assumptions to be satisfied, thus to achieve a good regret; given some environment, one should select the parameters of HOO such that the near-optimality dimension of the mean-payoff function is minimized. In the lack of knowledge of the mean-payoff function This might be difficult to achieve. Thus, ideally, these parameters should be selected adaptively based on the observation of some preliminary sample. For now, the investigation of this possibility is left for future work.*

4.3. Improving the running time when the time horizon is known. A deficiency of the basic HOO algorithm is that its computational complexity scales quadratically with the number of time steps. In this section we propose a simple modification to HOO that achieves essentially the same regret as HOO and whose computational complexity scales only log-linearly with the number of time steps. The needed amount of memory is still linear. We work out the case when the time horizon, n_0, is known in advance. The case of unknown horizon can be dealt with by resorting to the doubling trick.

We consider two modifications to the algorithm described in Section 3. First, the quantities $U_{h,i}(n)$ of (4.2) are redefined by replacing the factor $\ln n$ by $\ln n_0$, that is, now

$$U_{h,i}(n) = \widehat{\mu}_{h,i}(n) + \sqrt{\frac{2\ln n_0}{T_{h,i}(n)}} + \nu_1 \rho^h\,.$$

(This results in a policy exploring somewhat more uniformly the arms.) The definition of the B–values in terms of the $U_{h,i}(n)$ is unchanged. A pleasant consequence of the above modification is that the B–value of a given node changes only when this node is part of a path selected by the algorithm. Thus at each round n, only the nodes along the chosen path need to be updated according to the obtained reward.

However, and this is the reason for the second modification, in the basic algorithm, a path at round n may be of length linear in n (because the tree could have a depth linear in n). This is why we also truncate the trees $\mathcal{T}_n$ at a depth D_{n_0} of the order of $\ln n_0$. More precisely, the algorithm now selects the node (H_n, I_n) to pull at round n by following a path in the tree $\mathcal{T}_{n-1}$, starting from the root and choosing at each node the child with the highest B–value (with the new definition above using $\ln n_0$), and stopping either when it encounters a node which has not been expanded before or a node at depth equal to

$$D_{n_0} = \left\lceil \frac{(\ln n_0)/2 - \ln(1/\nu_1)}{\ln(1/\rho)} \right\rceil\,.$$

(It is assumed that $n_0 > 1/\nu_1^2$ so that $D_{n_0} \geqslant 1$.) Note that since no child of a node (D_{n_0}, i) located at depth D_{n_0} will ever be explored, its B–value at round $n \leqslant n_0$ simply equals $U_{D_{n_0},i}(n)$.

We call this modified version of HOO the *truncated HOO* algorithm. The computational complexity of updating all B–values at each round n is of the order of D_{n_0} and thus of the order of $\ln n_0$. The total computational complexity up to round n_0 is therefore of the order of $n_0 \ln n_0$, as claimed in the introduction of this section.

As the next theorem indicates this new procedure enjoys almost the same cumulative regret bound as the basic HOO algorithm.

THEOREM 4.2 (Upper bound on the regret of truncated HOO). *Fix a horizon n_0 such that $D_{n_0} \geqslant 1$. Then, the regret bound of Theorem 4.1 still holds true at round n_0 for truncated HOO up to an additional additive $4\sqrt{n_0}$ factor.*

4.4. Local assumptions. In this section we weaken somewhat the weak Lipschitz assumption and require it only to hold locally around the maximum. For the sake of simplicity and to derive exact constants we also state in a more explicit way the assumption on the near-optimality dimension. We then propose a simple and efficient adaptation of the HOO algorithm suited for this context.

4.4.1. *Modified set of assumptions.*

Assumptions Given the parameters of (the adaption of) HOO, that is, the real numbers $\nu_1 > 0$ and $\rho \in (0,1)$ and the tree of coverings $(\mathcal{P}_{h,i})$, there exists a dissimilarity function ℓ such that Assumption A1 (for some $\nu_2 > 0$) as well as the following two assumptions hold.

A2'. There exists $\varepsilon_0 > 0$ such that for all optimal subsets $\mathcal{A} \subset \mathcal{X}$ (i.e., $\sup_{x \in \mathcal{A}} f(x) = f^*$) with diameter $\operatorname{diam}(\mathcal{A}) \leqslant \varepsilon_0$,

$$f^* - \inf_{x \in \mathcal{A}} f(x) \leqslant \operatorname{diam}(\mathcal{A}) \,.$$

Further, there exists $L > 0$ such that for all $x \in \mathcal{X}_{\varepsilon_0}$ and $\varepsilon \in [0, \varepsilon_0]$,

$$\mathcal{B}\big(x,\ f^* - f(x) + \varepsilon\big) \ \subset \ \mathcal{X}_{L\left(2(f^* - f(x)) + \varepsilon\right)} \,.$$

A3. There exist $C > 0$ and $d > 0$ such that for all $\varepsilon \leqslant \varepsilon_0$,

$$\mathcal{N}\big(\mathcal{X}_{c\varepsilon},\, \ell,\, \varepsilon\big) \leqslant C\varepsilon^{-d} \,,$$

where $c = 4L\nu_1/\nu_2$.

When f satisfies Assumption A2', we say that f is ε_0–locally L–*weakly Lipschitz* w.r.t. ℓ. Note that this assumption was obtained by weakening the characterizations (4.4) and (4.5) of weak Lipschitzness.

Assumption A3 is not a real assumption but merely a reformulation of the definition of near optimality (with the small added ingredient that the limit can be achieved, see the second step of the proof of Theorem 4.1 in Section 6.1).

EXAMPLE 4.4. *We consider again the domain $\mathcal{X}$ and function f studied in Example 4.3 and prove (as announced beforehand) that f is ε_0–locally 2^{a-1}–weakly Lipschitz w.r.t. the dissimilarity ℓ defined by $\ell(x,y) = \|x - y\|^a$; which, in fact, holds for all ε_0.*

PROOF. Note that $x^* = (0, \dots, 0)$ is such that $f^* = 1 = f(x^*)$. Therefore, for all $x \in \mathcal{X}$,

$$f^* - f(x) = \|x\|^a = \ell(x^*, x) \,,$$

which yields the first part of Assumption A2'. To prove that the second part is true for $L = 2^{a-1}$ and with no constraint on the considered ε, we first note that since $a \geqslant 1$, it

holds by convexity that $(u+v)^a \leqslant 2^{a-1}(u^a+v^a)$ for all $u, v \in \mathbb{R}$. Now, for all $\varepsilon \geqslant 0$ and $y \in \mathcal{B}\big(x,\, \|x\|^a + \varepsilon\big)$, i.e., y such that $\ell(x,y) = \|x-y\|^a \leqslant \|x\|^a + \varepsilon$,

$$f^* - f(y) \;=\; \|y\|^a \;\leqslant\; \big(\|x\| + \|x-y\|\big)^a \;\leqslant\; 2^{a-1}\big(\|x\|^a + \|x-y\|^a\big) \;\leqslant\; 2^{a-1}\big(2\|x\|^a + \varepsilon\big)\,,$$

which concludes the proof of the second part of A2'. □

4.4.2. *Modified HOO algorithm.* We now describe the proposed modifications to the basic HOO algorithm.

We first consider, as a building block, the algorithm called *z–HOO*, which takes an integer z as an additional parameter to the ones of HOO. Algorithm z–HOO works as follows: it never plays any node with depth smaller or equal to $z-1$ and starts directly the selection of a new node at depth z. To do so, it first picks the node at depth z with the best B–value, chooses a path and then proceeds as the basic HOO algorithm. Note in particular that the initialization of this algorithm consists (in the first 2^z rounds) in playing once each of the 2^z nodes located at depth z in the tree (since by definition a node that has not been played yet has a B–value equal to $+\infty$). We note in passing that when $z=0$, algorithm z–HOO coincides with the basic HOO algorithm.

Algorithm *local HOO* employs the doubling trick in conjunction with consecutive instances of z–HOO. It works as follows. The integers $r \geqslant 1$ will index different regimes. The r–th regime starts at round $2^r - 1$ and ends when the next regime starts; it thus lasts for 2^r rounds. At the beginning of regime r, a fresh copy of z_r–HOO, where $z_r = \lceil \log_2 r \rceil$, is initialized and is then used throughout the regime.

Note that each fresh start needs to pull at least once each of the 2^{z_r} nodes located at depth z_r (the number of these nodes is $\approx r$). However, since round r lasts for 2^r time steps (which is exponentially larger than the number of nodes to explore), the time spent on the initialization of z_r–HOO in any regime r is greatly outnumbered by the time spent in the rest of the regime.

In the rest of this section, we propose first an upper bound on the regret of z–HOO (with exact and explicit constants). This result will play a key role in proving a bound on the performance of local HOO.

4.4.3. *Adaptation of the regret bound.* In the following we write h_0 for the smallest integer such that

$$2\nu_1 \rho^{h_0} < \varepsilon_0$$

and consider the algorithm z–HOO, where $z \geqslant h_0$. In particular, when $z = 0$ is chosen, the obtained bound is the same as the one of Theorem 4.1, except that the constants are given in analytic forms.

THEOREM 4.3 (Regret bound for z–HOO). *Consider z–HOO tuned with parameters ν_1 and ρ such that Assumptions A1, A2' and A3 hold for some dissimilarity ℓ and the values ν_2, L, ε_0, C, d. If, in addition, $z \geqslant h_0$ and $n \geqslant 2$ is large enough so that*

$$z \leqslant \frac{1}{d+2}\,\frac{\ln(4L\nu_1 n) - \ln(\gamma \ln n)}{\ln(1/\rho)}\,,$$

where

$$\gamma = \frac{4\,CL\nu_1 \nu_2^{-d}}{(1/\rho)^{d+1} - 1}\left(\frac{16}{\nu_1^2 \rho^2} + 9\right),$$

then the following bound holds for the expected regret of z–HOO:

$$\mathbb{E}\big[R_n\big] \leqslant \left(1 + \frac{1}{\rho^{d+2}}\right)\big(4L\nu_1 n\big)^{(d+1)/(d+2)}(\gamma \ln n)^{1/(d+2)} + \big(2^z - 1\big)\left(\frac{8\,\ln n}{\nu_1^2 \rho^{2z}} + 4\right).$$

The proof, which is a modification of the proof to Theorem 4.1, can be found in Section 6.3 of the Appendix. The main complication arises because the weakened assumptions do not allow one to reason about the smoothness at an arbitrary scale; this is essentially due to the threshold ε_0 used in the formulation of the assumptions. This is why in the proposed variant of HOO we discard nodes located too close to the root (at depth smaller than $h_0 - 1$). Note that in the bound the second term arises from playing in regions corresponding to the descendants of "poor" nodes located at level z. In particular, this term disappears when $z = 0$, in which case we get a bound on the regret of HOO provided that $2\nu_1 < \varepsilon_0$ holds.

EXAMPLE 4.5. *We consider again the setting of Examples 4.2, 4.3, and 4.4. The domain is $\mathcal{X} = [0,1]^D$ and the mean-payoff function f is defined by $f(x) = 1 - \|x\|_\infty^2$. We assume that HOO is run with parameters $\rho = (1/4)^{1/D}$ and $\nu_1 = 4$. We already proved that Assumptions A1, A2' and A3 are satisfied with the dissimilarity $\ell(x,y) = \|x - y\|_\infty^2$, the constants $\nu_2 = 1/4$, $L = 2$, $d = 0$, and[4] $C = 128^{D/2}$, as well as any $\varepsilon_0 > 0$ (that is, with $h_0 = 0$). Thus, resorting to Theorem 4.3 (applied with $z = 0$), we obtain*

$$\gamma = \frac{32 \times 128^{D/2}}{4^{1/D} - 1}\big(4^{2/D} + 9\big)$$

and get

$$\mathbb{E}\big[R_n\big] \leqslant \big(1 + 4^{2/D}\big)\sqrt{32\gamma\, n \ln n} = \sqrt{\exp\big(O(D)\big)\, n \ln n}\,.$$

Under the prescribed assumptions, the rate of convergence is of order $\sqrt{n}$ no matter the ambient dimension D. Although the rate is independent of D, the latter impacts the performance through the multiplicative factor in front of the rate, which is exponential in D. This is, however, not an artifact of our analysis, since it is natural that exploration in a D–dimensional space comes at a cost exponential in D. (The exploration performed by HOO combines an initial global search, which is bound to be exponential in D, and a local optimization, whose regret is of the order of $\sqrt{n}$.)

THEOREM 4.4 (Regret bound for local HOO). *Consider local HOO and assume that its parameters are tuned such that Assumptions A1, A2' and A3 hold for some dissimilarity ℓ. Then the expected regret of local HOO is bounded (in a distribution-dependent sense) as follows,*

$$\mathbb{E}\big[R_n\big] = \widetilde{O}\Big(n^{(d+1)/(d+2)}\Big)\,.$$

4.5. Minimax optimality in metric spaces. In this section we provide two theorems showing the minimax optimality of HOO in metric spaces. The notion of packing dimension is key.

DEFINITION 4.4 (Packing dimension). *The ℓ–packing dimension of a set $\mathcal{X}$ (w.r.t. a dissimilarity ℓ) is defined as*

$$\limsup_{\varepsilon \to 0} \; \frac{\ln \mathcal{N}(\mathcal{X}, \ell, \varepsilon)}{\ln(\varepsilon^{-1})}\,.$$

For instance, it is easy to see that whenever ℓ is a norm, compact subsets of $\mathbb{R}^D$ with non-empty interiors have a packing dimension of D.

Let $\mathcal{F}_{\mathcal{X},\ell}$ be the class of all bandit environments on $\mathcal{X}$ with a weak Lipschitz mean-payoff function (i.e., satisfying Assumption A2). For the sake of clarity, we now denote, for a bandit

[4]To compute C, one can first note that $4L\nu_1/\nu_2 = 128$; the question at hand for Assumption A3 to be satisfied is therefore to upper bound the number of balls of radius $\sqrt{\varepsilon}$ (w.r.t. the supremum norm $\|\cdot\|_\infty$) that can be packed in a ball of radius $\sqrt{128\varepsilon}$, giving rise to the bound $C \leqslant \sqrt{128}^D$.

strategy φ and a bandit environment M on $\mathcal{X}$, the expectation of the cumulative regret of φ over M at time n by $\mathbb{E}_M\big[R_n(\varphi)\big]$.

The following theorem provides a uniform upper bound on the regret of HOO over this class of environments. (The constant γ appearing in the statement depends only on $\mathcal{X}$, ν_1, ρ, ℓ, ν_2, D'.)

THEOREM 4.5 (Uniform upper bound on the regret of HOO). *Assume that $\mathcal{X}$ has a finite ℓ–packing dimension D and that the parameters of HOO are such that A1 is satisfied. Then, for all $D' > D$ there exists a constant γ such that for all $n \geqslant 1$,*

$$\sup_{M \in \mathcal{F}_{\mathcal{X},\ell}} \mathbb{E}_M\big[R_n(\text{HOO})\big] \leqslant \gamma\, n^{(D'+1)/(D'+2)} \left(\ln n\right)^{1/(D'+2)}.$$

The next result shows that in the case of metric spaces this upper bound is optimal up to a multiplicative logarithmic factor. Note that if $\mathcal{X}$ is a large enough compact subset of $\mathbb{R}^D$ with non-empty interior and the dissimilarity ℓ is some norm of $\mathbb{R}^D$, then the assumption of the following theorem is satisfied.

THEOREM 4.6 (Uniform lower bound). *Consider a set $\mathcal{X}$ equipped with a dissimilarity ℓ that is a metric. Assume that there exists some constant $c \in (0,1]$ such that for all $\varepsilon \leqslant 1$, the packing numbers satisfy $\mathcal{N}(\mathcal{X},\ell,\varepsilon) \geqslant c\,\varepsilon^{-D} \geqslant 2$. Then, there exist two constants $N(c,D)$ and $\gamma(c,D)$ depending only on c and D such that for all bandit strategies φ and all $n \geqslant N(c,D)$,*

$$\sup_{M \in \mathcal{F}_{\mathcal{X},\ell}} \mathbb{E}_M\big[R_n(\varphi)\big] \geqslant \gamma(c,D)\; n^{(D+1)/(D+2)}.$$

The reader interested in the explicit expressions of $N(c,D)$ and $\gamma(c,D)$ is referred to the last lines of the proof of the theorem in the Appendix.

5. Discussion

In this section we would like to shed some light on the results of the previous sections. In particular we generalize the situation of Example 4.5, discuss the regret that we can obtain, and compare it with what could be obtained by previous works.

5.1. Example. Equip $\mathcal{X} = [0,1]^D$ with a norm $\|\cdot\|$ and assume that the mean-payoff function f is locally equivalent to a Hölder continuous function with degree $\alpha \in [0,\infty)$ around any global maximum x^* of f (these maxima being in addition assumed to be in finite number); that is,

$$f(x^*) - f(x) = \Theta\big(\|x - x^*\|^\alpha\big) \qquad \text{as} \quad x \to x^*.$$

This means that there exist $c_1, c_2, \delta > 0$ such that for all x satisfying $\|x - x^*\| \leqslant \delta$,

$$c_2 \|x - x^*\|^\alpha \leqslant f(x^*) - f(x) \leqslant c_1 \|x - x^*\|^\alpha.$$

In particular, one can check that Assumption A2' is satisfied for the dissimilarity defined by $\ell_{c,\beta}(x,y) = c\|x-y\|^\beta$, where $\beta \leqslant \alpha$ (and $c \geqslant c_1$ when $\beta = \alpha$). We further assume that HOO is run with parameters ν_1 and ρ and a tree of dyadic partitions such that Assumption A1 is satisfied as well (see Examples 4.1 and 4.2 for explicit values of these parameters in the case of the Euclidean or the supremum norms over the unit cube). The following statements can then be formulated on the expected regret of HOO.

- **Known smoothness:** If we know the true smoothness of f around its maxima, then we set $\beta = \alpha$ and $c \geqslant c_1$. This choice $\ell_{c_1,\alpha}$ of a dissimilarity is such that f is locally weak-Lipschitz with respect to it and the near-optimality dimension is $d = 0$ (cf. Example 4.3). Theorem 4.4 thus implies that the expected regret of local HOO is $\widetilde{O}(\sqrt{n})$, i.e., *the rate of the bound is independent of the dimension D.*

- **Smoothness underestimated:** Here, we assume that the true smoothness of f around its maxima is unknown and that it is underestimated by choosing $\beta < \alpha$ (and some c). Then f is still locally weak-Lipschitz with respect to the dissimilarity $\ell_{c,\beta}$ and the near-optimality dimension is $d = D(1/\beta - 1/\alpha)$, as shown in Example 4.3; the regret of HOO is $\widetilde{O}\big(n^{(d+1)/(d+2)}\big)$.
- **Smoothness overestimated:** Now, if the true smoothness is overestimated by choosing $\beta > \alpha$ or $\alpha = \beta$ and $c < c_1$, then the assumption of weak Lipschitzness is violated and we are unable to provide any guarantee on the behavior of HOO. The latter, when used with an overestimated smoothness parameter, may lack exploration and exploit too heavily from the beginning. As a consequence, it may get stuck in some local optimum of f, missing the global one(s) for a very long time (possibly indefinitely). Such a behavior is illustrated in the example provided in Coquelin and Munos [2007] and showing the possible problematic behavior of the closely related algorithm UCT of Kocsis and Szepesvari [2006]. UCT is an example of an algorithm overestimating the smoothness of the function; this is because the B–values of UCT are defined similarly to the ones of the HOO algorithm but without the third term in the definition (4.2) of the U–values. This corresponds to an assumed infinite degree of smoothness (that is, to a locally constant mean-payoff function).

5.2. Relation to previous works. Several works [Agrawal, 1995b, Kleinberg, 2004, Cope, 2004, Auer et al., 2007, Kleinberg et al., 2008a] have considered continuum-armed bandits in Euclidean or, more generally, normed or metric spaces and provided upper and lower bounds on the regret for given classes of environments.

- Cope [2004] derived a $\widetilde{O}(\sqrt{n})$ bound on the regret for compact and convex subsets of $\mathbb{R}^d$ and mean-payoff functions with a unique minimum and second-order smoothness.
- Kleinberg [2004] considered mean-payoff functions f on the real line that are Hölder continuous with degree $0 < \alpha \leqslant 1$. The derived regret bound is $\Theta\big(n^{(\alpha+1)/(\alpha+2)}\big)$.
- Auer et al. [2007] extended the analysis to classes of functions with only a local Hölder assumption around the maxima, where the allowed smoothness degree is also larger: $\alpha \in [0, \infty)$. They derived the regret bound

$$\Theta\Big(n^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}}\Big)\,,$$

  where the parameter β is such that the Lebesgue measure of ε–optimal arm is $O(\varepsilon^{\beta})$.
- Another setting is the one of Kleinberg et al. [2008a], who considered a metric space $(\mathcal{X}, \ell)$ and assumed that f is Lipschitz w.r.t. ℓ. The obtained regret bound is $\widetilde{O}\big(n^{(d+1)/(d+2)}\big)$, where d is the *zooming dimension*. The latter is defined similarly to our near-optimality dimension, except firstly, that covering numbers instead of packing numbers are used and secondly, that sets of the form $\mathcal{X}_{\varepsilon} \setminus \mathcal{X}_{\varepsilon/2}$ are considered instead of the $\mathcal{X}_{c\varepsilon}$. When $(\mathcal{X}, \ell)$ is a metric space, covering and packing numbers are within a constant factor to each other, and therefore, one may prove that the zooming and near-optimality dimensions are also equal.

Our main contribution compared to Kleinberg et al. [2008a] is that our weak Lipschitz assumption, which is substantially weaker than the global Lipschitz condition imposed in Kleinberg et al. [2008a], enables our algorithm to work better in some common situations; for instance, when the mean-payoff function is locally smoothn with a smoothness order larger than 1.

For an illustration, consider again the example of Section 5.1. The result of Auer et al. [2007] shows that for $D = 1$, the regret is $\Theta(\sqrt{n})$ (since here $\beta = 1/\alpha$, with the notation above). Our result extends the $\sqrt{n}$ rate of the regret bound to any dimension D.

Now, we compare this result with the bounds that result from Kleinberg et al. [2008a]. The case when $\alpha \leqslant 1$ lead to identical rates; however, this is no longer the case when $\alpha > 1$. Multiple issues arise indeed in the latter case. First, the analysis of Kleinberg et al. [2008a] assumes globally Lipschitz functions with respect to the chosen metric. Now, a function that is globally Lipschitz w.r.t. ℓ_α is constant; furthermore, in this case, ℓ_α is not even a metric, while the results of Kleinberg et al. [2008a] rely crucially on the use of a metric.

To avoid these issues one may attempt a fall-back to (say) the Euclidean metric, so that the requirement that f is Lipschitz with respect to the metric is satisfied. The zooming dimension then becomes $(1 - 1/\alpha)D$ (see Example 4.3), while the regret bound of Kleinberg et al. becomes $\widetilde{O}\big(n^{(D(\alpha-1)+\alpha)/(D(\alpha-1)+2\alpha)}\big)$. Note that this regret is strictly worse than $\widetilde{O}(\sqrt{n})$ and in fact becomes closer to the slow rate $\widetilde{O}\big(n^{(D+1)/(D+2)}\big)$ as $\alpha \to \infty$.

6. Proofs

6.1. Proof of Theorem 4.1 (main upper bound on the regret of HOO). We begin with three lemmas. The proofs of Lemmas 4.3 and 4.4 rely on concentration-of-measure techniques, while the one of Lemma 4.2 follows from a simple case study. Let us fix some path $(0, 1)$, $(1, i_1^*)$, $(2, i_2^*)$, ... of optimal nodes, starting from the root. That is, denoting $i_0^* = 1$, we mean that for all $j \geqslant 1$, the suboptimality of (j, i_j^*) equals $\Delta_{j,i_j^*} = 0$ and (j, i_j^*) is a child of $(j - 1, i_{j-1}^*)$.

LEMMA 4.2. *Let (h, i) be a suboptimal node. Let $0 \leqslant k \leqslant h - 1$ be the largest depth such that (k, i_k^*) is on the path from the root $(0, 1)$ to (h, i). Then for all integers $u \geqslant 0$, we have*

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant u + \sum_{t=u+1}^{n} \mathbb{P}\Big\{ \big[U_{s,i_s^*}(t) \leqslant f^* \text{ for some } s \in \{k+1, \dots, t-1\}\big] \\ \text{or} \quad \big[T_{h,i}(t) > u \ \text{ and } \ U_{h,i}(t) > f^*\big] \Big\}.$$

PROOF. Consider a given round $t \in \{1, \dots, n\}$. If $(H_t, I_t) \in \mathcal{C}(h, i)$, then this is because the child $(k + 1, i')$ of (k, i_k^*) on the path to (h, i) had a better B–value than its brother $(k + 1, i_{k+1}^*)$. Since by definition, B–values can only increase on a chosen path, this entails that $B_{k+1,i_{k+1}^*} \leqslant B_{k+1,i'}(t) \leqslant B_{h,i}(t)$. This is turns implies, again by definition of the B–values, that $B_{k+1,i_{k+1}^*}(t) \leqslant U_{h,i}(t)$. Thus,

$$\big\{(H_t, I_t) \in \mathcal{C}(h, i)\big\} \subset \big\{U_{h,i}(t) \geqslant B_{k+1,i_{k+1}^*}(t)\big\} \subset \big\{U_{h,i}(t) > f^*\big\} \cup \big\{B_{k+1,i_{k+1}^*}(t) \leqslant f^*\big\}.$$

But, once again by definition of B–values,

$$\big\{B_{k+1,i_{k+1}^*}(t) \leqslant f^*\big\} \subset \big\{U_{k+1,i_{k+1}^*}(t) \leqslant f^*\big\} \cup \big\{B_{k+2,i_{k+2}^*}(t) \leqslant f^*\big\},$$

and the argument can be iterated. Since up to round t no more than t nodes have been played (including the suboptimal node (h, i)), we know that (t, i_t^*) has not been played so far and thus has a B–value equal to $+\infty$. (Some of the previous optimal nodes could also have had an infinite U–value, if not played so far.) We thus have proved the inclusion

(4.6)

$$\big\{(H_t, I_t) \in \mathcal{C}(h, i)\big\} \subset \big\{U_{h,i}(t) > f^*\big\} \cup \Big(\big\{U_{k+1,i_{k+1}^*}(t) \leqslant f^*\big\} \cup \ldots \cup \big\{U_{t-1,i_{t-1}^*}(t) \leqslant f^*\big\}\Big).$$

Now, for any integer $u \geqslant 0$ it holds that

$$\begin{aligned} T_{h,i}(n) &= \sum_{t=1}^{n} \mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i),\ T_{h,i}(t)\leqslant u\}} + \sum_{t=1}^{n} \mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i),\ T_{h,i}(t)>u\}} \\ &\leqslant u + \sum_{t=u+1}^{n} \mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i),\ T_{h,i}(t)>u\}}\,, \end{aligned}$$

where we used for the inequality the fact that the quantities $T_{h,i}(t)$ are constant from t to $t+1$, except when $(H_t, I_t) \in \mathcal{C}(h,i)$, in which case, they increase by 1; therefore, on the one hand, at most u of the $T_{h,i}(t)$ can be smaller than u and on the other hand, $T_{h,i}(t) > u$ can only happen if $t > u$. Using (4.6) and then taking expectations yields the result. □

LEMMA 4.3. *Let Assumptions A1 and A2 hold. Then, for all optimal nodes (h,i) and for all integers $n \geqslant 1$,*

$$\mathbb{P}\big\{U_{h,i}(n) \leqslant f^*\big\} \leqslant n^{-3}\,.$$

PROOF. On the event that (h,i) was not played during the first n rounds, one has, by convention, $U_{h,i}(n) = +\infty$. In the sequel, we therefore restrict our attention to the event $\big\{T_{h,i}(n) \geqslant 1\big\}$.

Lemma 4.1 with $c = 0$ ensures that $f^* - f(x) \leqslant \nu_1 \rho^h$ for all arms $x \in \mathcal{P}_{h,i}$. Hence,

$$\sum_{t=1}^{n} \big(f(X_t) + \nu_1\rho^h - f^*\big)\, \mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i)\}} \geqslant 0$$

and therefore,

$$\begin{aligned} &\mathbb{P}\big\{U_{h,i}(n) \leqslant f^* \quad \text{and} \quad T_{h,i}(n) \geqslant 1\big\} \\ &\quad= \mathbb{P}\left\{\widehat{\mu}_{h,i}(n) + \sqrt{\frac{2\ln n}{T_{h,i}(n)}} + \nu_1\rho^h \leqslant f^* \quad \text{and} \quad T_{h,i}(n) \geqslant 1\right\} \\ &\quad= \mathbb{P}\left\{T_{h,i}(n)\,\widehat{\mu}_{h,i}(n) + T_{h,i}(n)\,\big(\nu_1\rho^h - f^*\big) \leqslant -\sqrt{2\,T_{h,i}(n)\ln n} \quad \text{and} \quad T_{h,i}(n) \geqslant 1\right\} \\ &\quad\leqslant \mathbb{P}\Bigg\{\sum_{t=1}^{n}\big(Y_t - f(X_t)\big)\mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i)\}} + \sum_{t=1}^{n}\big(f(X_t) + \nu_1\rho^h - f^*\big)\mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i)\}} \\ &\qquad\qquad \leqslant -\sqrt{2\,T_{h,i}(n)\ln n} \quad \text{and} \quad T_{h,i}(n) \geqslant 1\Bigg\} \\ &\quad\leqslant \mathbb{P}\left\{\sum_{t=1}^{n}\big(f(X_t) - Y_t\big)\mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i)\}} \geqslant \sqrt{2\,T_{h,i}(n)\ln n} \quad \text{and} \quad T_{h,i}(n) \geqslant 1\right\}. \end{aligned}$$

We take care of the last term with a union bound and the Hoeffding-Azuma inequality for martingale differences, see Theorem 10.1.

To do this in a rigorous manner, we need to define a sequence of (random) stopping times when arms in $\mathcal{C}(h,i)$ were pulled:

$$T_j = \min\big\{t :\ T_{h,i}(t) = j\big\}\,, \quad j = 1, 2, \ldots\,.$$

Note that $1 \leqslant T_1 < T_2 < \ldots$, hence it holds that $T_j \geqslant j$. We denote by $\widetilde{X}_j = X_{T_j}$ the j–th arm pulled in the region corresponding to $\mathcal{C}(h,i)$. Its associated corresponding reward equals $\widetilde{Y}_j = Y_{T_j}$ and

$$\mathbb{P}\left\{\sum_{t=1}^{n}\big(f(X_t) - Y_t\big)\mathbb{I}_{\{(H_t,I_t)\in\mathcal{C}(h,i)\}} \geqslant \sqrt{2\,T_{h,i}(n)\ln n} \quad \text{and} \quad T_{h,i}(n) \geqslant 1\right\}$$

$$
\begin{aligned}
&= \mathbb{P}\left\{ \sum_{j=1}^{T_{h,i}(n)} \Big(f\big(\widetilde{X}_j\big) - \widetilde{Y}_j \Big) \geqslant \sqrt{2\, T_{h,i}(n) \ln n} \quad \text{and} \quad T_{h,i}(n) \geqslant 1 \right\} \\
&\leqslant \sum_{t=1}^{n} \mathbb{P}\left\{ \sum_{j=1}^{t} \Big(f\big(\widetilde{X}_j\big) - \widetilde{Y}_j \Big) \geqslant \sqrt{2\, t \ln n} \right\},
\end{aligned}
$$

where we used a union bound to get the last inequality.

We claim that

$$
Z_t = \sum_{j=1}^{t} \Big(f(\widetilde{X}_j) - \widetilde{Y}_j \Big)
$$

is a martingale (with respect to the filtration it generates). This follows, via optional skipping (see [Doob, 1953, Chapter VII, Theorem 2.3]), from the facts that

$$
\sum_{t=1}^{n} \big(f(X_t) - Y_t \big) \mathbb{I}_{\{(H_t, I_t) \in \mathcal{C}(h,i)\}}
$$

is a martingale w.r.t. the filtration $\mathcal{F}_t = \sigma(X_1, Y_1, \ldots, X_t, Y_t)$ and that the events $\{T_j = k\} \in \mathcal{F}_{k-1}$ for all $k \geqslant j$.

Applying the Hoeffding-Azuma inequality for martingale differences (see Theorem 10.1), using the boundedness of the ranges of the induced martingale difference sequence, we then get, for each $t \geqslant 1$,

$$
\mathbb{P}\left\{ \sum_{j=1}^{t} \Big(f\big(\widetilde{X}_j\big) - \widetilde{Y}_j \Big) \geqslant \sqrt{2\, t \ln n} \right\} \leqslant \exp\left(- \frac{2 \left(\sqrt{2\, t \ln n} \right)^2}{t} \right) = n^{-4},
$$

which concludes the proof. □

LEMMA 4.4. *For all integers $t \leqslant n$, for all suboptimal nodes (h, i) such that $\Delta_{h,i} > \nu_1 \rho^h$, and for all integers $u \geqslant 1$ such that*

$$
u \geqslant \frac{8 \ln n}{(\Delta_{h,i} - \nu_1 \rho^h)^2},
$$

one has

$$
\mathbb{P}\big\{ U_{h,i}(t) > f^* \ \text{ and } \ T_{h,i}(t) > u \big\} \leqslant t\, n^{-4}.
$$

PROOF. The u mentioned in the statement of the lemma are such that

$$
\frac{\Delta_{h,i} - \nu_1 \rho^h}{2} \geqslant \sqrt{\frac{2 \ln n}{u}}, \qquad \text{thus} \qquad \sqrt{\frac{2 \ln t}{u}} + \nu_1 \rho^h \leqslant \frac{\Delta_{h,i} + \nu_1 \rho^h}{2}.
$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}\big\{ U_{h,i}(t) > f^* \quad \text{and} \quad T_{h,i}(t) > u \big\} \\
&\quad = \mathbb{P}\left\{ \widehat{\mu}_{h,i}(t) + \sqrt{\frac{2 \ln t}{T_{h,i}(t)}} + \nu_1 \rho^h > f^*_{h,i} + \Delta_{h,i} \quad \text{and} \quad T_{h,i}(t) > u \right\} \\
&\quad \leqslant \mathbb{P}\left\{ \widehat{\mu}_{h,i}(t) > f^*_{h,i} + \frac{\Delta_{h,i} - \nu_1 \rho^h}{2} \quad \text{and} \quad T_{h,i}(t) > u \right\} \\
&\quad \leqslant \mathbb{P}\left\{ T_{h,i}(t) \left(\widehat{\mu}_{h,i}(t) - f^*_{h,i} \right) > \frac{\Delta_{h,i} - \nu_1 \rho^h}{2}\, u \quad \text{and} \quad T_{h,i}(t) > u \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{P}\left\{\sum_{s=1}^{t}\bigl(Y_s - f^*_{h,i}\bigr)\mathbb{I}_{\{(H_s,I_s)\in\mathcal{C}(h,i)\}} > \frac{\Delta_{h,i}-\nu_1\rho^h}{2}\,u \quad \text{and} \quad T_{h,i}(t) > u\right\} \\
&\leqslant \mathbb{P}\left\{\sum_{s=1}^{t}\bigl(Y_s - f(X_s)\bigr)\mathbb{I}_{\{(H_s,I_s)\in\mathcal{C}(h,i)\}} > \frac{\Delta_{h,i}-\nu_1\rho^h}{2}\,u \quad \text{and} \quad T_{h,i}(t) > u\right\}.
\end{aligned}
$$

Now it follows from the same arguments as in the proof of Lemma 4.3 (optional skipping, the Hoeffding-Azuma inequality, and a union bound) that

$$
\begin{aligned}
&\mathbb{P}\left\{\sum_{s=1}^{t}\bigl(Y_s - f(X_s)\bigr)\,\mathbb{I}_{\{(H_s,I_s)\in\mathcal{C}(h,i)\}} > \frac{\Delta_{h,i}-\nu_1\rho^h}{2}\,u \quad \text{and} \quad T_{h,i}(t) > u\right\} \\
&\quad\leqslant \sum_{s=u+1}^{t}\exp\left(-\frac{2}{s}\left(\frac{(\Delta_{h,i}-\nu_1\rho^h)\,u}{2}\right)^2\right) \leqslant t\,\exp\left(-\frac{1}{2}\,u\,\bigl(\Delta_{h,i}-\nu_1\rho^h\bigr)^2\right) \leqslant t\,n^{-4}\,,
\end{aligned}
$$

where we used the stated bound on u to obtain the last inequality. □

Combining the results of Lemmas 4.2, 4.3, and 4.4 leads to the following key result bounding the expected number of visits to descendants of a "poor" node.

LEMMA 4.5. *Under Assumptions A1 and A2, for all suboptimal nodes (h,i) with $\Delta_{h,i} > \nu_1\rho^h$, we have, for all $n \geqslant 1$,*

$$
\mathbb{E}[T_{h,i}(n)] \leqslant \frac{8\,\ln n}{(\Delta_{h,i}-\nu_1\rho^h)^2} + 4\,.
$$

PROOF. We take u as the upper integer part of $(8\,\ln n)/(\Delta_{h,i}-\nu_1\rho^h)^2$ and use union bounds to get from Lemma 4.2 the bound

$$
\begin{aligned}
\mathbb{E}\bigl[T_{h,i}(n)\bigr] \leqslant{}& \frac{8\,\ln n}{(\Delta_{h,i}-\nu_1\rho^h)^2} + 1 \\
&+ \sum_{t=u+1}^{n}\left(\mathbb{P}\bigl\{T_{h,i}(t) > u \ \text{ and } \ U_{h,i}(t) > f^*\bigr\} + \sum_{s=1}^{t-1}\mathbb{P}\bigl\{U_{s,i^*_s}(t) \leqslant f^*\bigr\}\right).
\end{aligned}
$$

Lemmas 4.3 and 4.4 further bound the quantity of interest as

$$
\mathbb{E}\bigl[T_{h,i}(n)\bigr] \leqslant \frac{8\,\ln n}{(\Delta_{h,i}-\nu_1\rho^h)^2} + 1 + \sum_{t=u+1}^{n}\left(t\,n^{-4} + \sum_{s=1}^{t-1}t^{-3}\right)
$$

and we now use the crude upper bounds

$$
1 + \sum_{t=u+1}^{n}\left(t\,n^{-4} + \sum_{s=1}^{t-1}t^{-3}\right) \leqslant 1 + \sum_{t=1}^{n}\bigl(n^{-3}+t^{-2}\bigr) \leqslant 2 + \pi^2/6 \leqslant 4
$$

to get the proposed statement. □

PROOF. **(of Theorem 4.1)** First, let us fix $d' > d$. The statement will be proven in four steps.

First step. For all $h = 0, 1, 2, \ldots$, denote by $\mathcal{I}_h$ the set of those nodes at depth h that are $2\nu_1\rho^h$–optimal, i.e., the nodes (h,i) such that $f^*_{h,i} \geqslant f^* - 2\nu_1\rho^h$. (Of course, $\mathcal{I}_0 = \{(0,1)\}$.) Then, let $\mathcal{I}$ be the union of these sets when h varies. Further, let $\mathcal{J}$ be the set of nodes that are not in $\mathcal{I}$ but whose parent is in $\mathcal{I}$. Finally, for $h = 1, 2, \ldots$ we denote by $\mathcal{J}_h$ the nodes in $\mathcal{J}$ that are located at depth h in the tree (i.e., whose parent is in $\mathcal{I}_{h-1}$).

Lemma 4.5 bounds in particular the expected number of times each node $(h,i) \in \mathcal{J}_h$ is visited. Since for these nodes $\Delta_{h,i} > 2\nu_1\rho^h$, we get

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant \frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\,.$$

Second step. We bound the cardinality $|\mathcal{I}_h|$ of $\mathcal{I}_h$. We start with the case $h \geqslant 1$. By definition, when $(h,i) \in \mathcal{I}_h$, one has $\Delta_{h,i} \leqslant 2\nu_1\rho^h$, so that by Lemma 4.1 the inclusion $\mathcal{P}_{h,i} \subset \mathcal{X}_{4\nu_1\rho^h}$ holds. Since by Assumption A1, the sets $\mathcal{P}_{h,i}$ contain disjoint balls of radius $\nu_2\rho^h$, we have that

$$|\mathcal{I}_h| \leqslant \mathcal{N}\big(\cup_{(h,i)\in\mathcal{I}_h}\mathcal{P}_{h,i},\,\ell,\,\nu_2\rho^h\big) \leqslant \mathcal{N}\big(\mathcal{X}_{4\nu_1\rho^h},\,\ell,\,\nu_2\rho^h\big) = \mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\nu_2\rho^h},\,\ell,\,\nu_2\rho^h\big)\,.$$

We prove below that there exists a constant C such that for all $\varepsilon \leqslant \nu_2$,

$$\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big) \leqslant C\,\varepsilon^{-d'}\,. \tag{4.7}$$

Thus we obtain the bound $|\mathcal{I}_h| \leqslant C\left(\nu_2\rho^h\right)^{-d'}$ for all $h \geqslant 1$. We note that the obtained bound $|\mathcal{I}_h| \leqslant C\left(\nu_2\rho^h\right)^{-d'}$ is still valid for $h = 0$, since $|\mathcal{I}_0| = 1$.

It only remains to prove (4.7). Since $d' > d$, where d is the near-optimality of f, we have, by definition, that

$$\limsup_{\varepsilon\to 0} \frac{\ln\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big)}{\ln\big(\varepsilon^{-1}\big)} \leqslant d\,,$$

and thus, there exists $\varepsilon_{d'} > 0$ such that for all $\varepsilon \leqslant \varepsilon_{d'}$,

$$\frac{\ln\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big)}{\ln\big(\varepsilon^{-1}\big)} \leqslant d'\,,$$

which in turn implies that for all $\varepsilon \leqslant \varepsilon_{d'}$,

$$\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big) \leqslant \varepsilon^{-d'}\,.$$

The result is proved with $C = 1$ if $\varepsilon_{d'} \geqslant \nu_2$. Now, consider the case $\varepsilon_{d'} < \nu_2$. Given the definition of packing numbers, it is straightforward that for all $\varepsilon \in \big[\varepsilon_{d'},\,\nu_2\big]$,

$$\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big) \leqslant u_{d'} \stackrel{\text{def}}{=} \mathcal{N}\big(\mathcal{X},\,\ell,\,\varepsilon_{d'}\big)\,;$$

therefore, for all $\varepsilon \in \big[\varepsilon_{d'},\,\nu_2\big]$,

$$\mathcal{N}\big(\mathcal{X}_{(4\nu_1/\nu_2)\,\varepsilon},\,\ell,\,\varepsilon\big) \leqslant u_{d'}\,\frac{\nu_2^{d'}}{\varepsilon^{d'}} = C\varepsilon^{-d'}$$

for the choice $C = \max\big\{1,\ u_{d'}\,\nu_2^{d'}\big\}$. Because we take the maximum with 1, the stated inequality also holds for $\varepsilon \leqslant \varepsilon^{-d'}$, which concludes the proof of (4.7).

Third step. Let $H \geqslant 1$ be an integer to be chosen later. We partition the nodes of the infinite tree $\mathcal{T}$ into three subsets, $\mathcal{T} = \mathcal{T}^1 \cup \mathcal{T}^2 \cup \mathcal{T}^3$, as follows. Let the set $\mathcal{T}^1$ contain the descendants of the nodes in $\mathcal{I}_H$ (by convention, a node is considered its own descendant, hence the nodes of $\mathcal{I}_H$ are included in $\mathcal{T}^1$); let $\mathcal{T}^2 = \cup_{0\leqslant h<H}\mathcal{I}_h$; and let $\mathcal{T}^3$ contain the descendants of the nodes in $\cup_{1\leqslant h\leqslant H}\mathcal{J}_h$. Thus, $\mathcal{T}^1$ and $\mathcal{T}^3$ are potentially infinite, while $\mathcal{T}^2$ is finite.

We recall that we denote by (H_t, I_t) the node that was chosen by HOO in round t. From the definition of the algorithm, each node is played at most once, thus no two such random variables are equal when t varies. We decompose the regret according to which of the sets $\mathcal{T}^j$ the nodes (H_t, I_t) belong to:

$$\mathbb{E}\big[R_n\big] = \mathbb{E}\left[\sum_{t=1}^{n}(f^* - f(X_t))\right] = \mathbb{E}\big[R_{n,1}\big] + \mathbb{E}\big[R_{n,2}\big] + \mathbb{E}\big[R_{n,3}\big]\,,$$

$$\text{where} \qquad R_{n,i} = \sum_{t=1}^{n} \bigl(f^* - f(X_t)\bigr) \mathbb{I}_{\{(H_t,I_t)\in\mathcal{T}^i\}}\,, \qquad \text{for } i = 1, 2, 3.$$

The contribution from $\mathcal{T}^1$ is easy to bound. By definition any node in $\mathcal{I}_H$ is $2\nu_1\rho^H$–optimal. Hence, by Lemma 4.1, the corresponding domain is included in $\mathcal{X}_{4\nu_1\rho^H}$. By definition of a tree of coverings, the domains of the descendants of these nodes are still included in $\mathcal{X}_{4\nu_1\rho^H}$. Therefore,

$$\mathbb{E}\bigl[R_{n,1}\bigr] \leqslant 4\nu_1\rho^H\, n\,.$$

For $h \geqslant 0$, consider a node $(h,i) \in \mathcal{T}^2$. It belongs to $\mathcal{I}_h$ and is therefore $2\nu_1\rho^h$–optimal. By Lemma 4.1, the corresponding domain is included in $\mathcal{X}_{4\nu_1\rho^h}$. By the result of the second step of this proof and using that each node is played at most once, one gets

$$\mathbb{E}\bigl[R_{n,2}\bigr] \leqslant \sum_{h=0}^{H-1} 4\nu_1\rho^h\,|\mathcal{I}_h| \leqslant 4C\nu_1\nu_2^{-d'} \sum_{h=0}^{H-1} \rho^{h(1-d')}\,.$$

We finish by bounding the contribution from $\mathcal{T}^3$. We first remark that since the parent of any element $(h,i) \in \mathcal{J}_h$ is in $\mathcal{I}_{h-1}$, by Lemma 4.1 again, we have that $\mathcal{P}_{h,i} \subset \mathcal{X}_{4\nu_1\rho^{h-1}}$. We now use the first step of this proof to get

$$\mathbb{E}\bigl[R_{n,3}\bigr] \leqslant \sum_{h=1}^{H} 4\nu_1\rho^{h-1} \sum_{i\,:\,(h,i)\in\mathcal{J}_h} \mathbb{E}\bigl[T_{h,i}(n)\bigr] \leqslant \sum_{h=1}^{H} 4\nu_1\rho^{h-1}\,|\mathcal{J}_h| \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right).$$

Now, it follows from the fact that the parent of $\mathcal{J}_h$ is in $\mathcal{I}_{h-1}$ that $|\mathcal{J}_h| \leqslant 2|\mathcal{I}_{h-1}|$ when $h \geqslant 1$. Substituting this and the bound on $|\mathcal{I}_{h-1}|$ obtained in the second step of this proof, we get

$$\begin{aligned}
\mathbb{E}\bigl[R_{n,3}\bigr] &\leqslant \sum_{h=1}^{H} 4\nu_1\rho^{h-1} \left(2C\left(\nu_2\rho^{h-1}\right)^{-d'}\right) \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right) \\
&\leqslant 8C\nu_1\nu_2^{-d'} \sum_{h=1}^{H} \rho^{h(1-d')+d'-1} \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right).
\end{aligned}$$

Fourth step. Putting the obtained bounds together, we get

$$\begin{aligned}
\mathbb{E}\bigl[R_n\bigr] &\leqslant 4\nu_1\rho^H\, n + 4C\nu_1\nu_2^{-d'} \sum_{h=0}^{H-1} \rho^{h(1-d')} + 8C\nu_1\nu_2^{-d'} \sum_{h=1}^{H} \rho^{h(1-d')+d'-1} \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right) \\
&= O\left(n\rho^H + (\ln n)\sum_{h=1}^{H} \rho^{-h(1+d')}\right) = O\Bigl(n\rho^H + \rho^{-H(1+d')}\ln n\Bigr) \qquad (4.8)
\end{aligned}$$

(recall that $\rho < 1$). Note that all constants hidden in the O symbol only depend on ν_1, ν_2, ρ and d'.

Now, by choosing H such that $\rho^{-H(d'+2)}$ is of the order of $n/\ln n$, that is, ρ^H is of the order of $(n/\ln n)^{-1/(d'+2)}$, we get the desired result, namely,

$$\mathbb{E}\bigl[R_n\bigr] = O\Bigl(n^{(d'+1)/(d'+2)}\,(\ln n)^{1/(d'+2)}\Bigr)\,.$$

□

6.2. Proof of Theorem 4.2 (regret bound for truncated HOO). The proof follows from an adaptation of the proof of Theorem 4.1 and of its associated lemmas; for the sake of clarity and precision, we explicitly state the adaptations of the latter.

Adaptations of the lemmas. Remember that D_{n_0} denotes the maximum depth of the tree, given horizon n_0. The adaptation of Lemma 4.2 is done as follows. Let (h,i) be a suboptimal node with $h \leqslant D_{n_0}$ and let $0 \leqslant k \leqslant h-1$ be the largest depth such that (k, i_k^*) is on the path from the root $(0,1)$ to (h,i). Then, for all integers $u \geqslant 0$, one has

$$\mathbb{E}\big[T_{h,i}(n_0)\big] \leqslant u + \sum_{t=u+1}^{n_0} \mathbb{P}\Big\{ \big[U_{s,i_s^*}(t) \leqslant f^* \text{ for some } s \text{ with } k+1 \leqslant s \leqslant \min\{D_{n_0}, n_0\}\big] \\ \text{or} \quad \big[T_{h,i}(t) > u \ \text{ and } \ U_{h,i}(t) > f^*\big] \Big\}.$$

As for Lemma 4.3, its straightforward adaptation states that under Assumptions A1 and A2, for all optimal nodes (h,i) with $h \leqslant D_{n_0}$ and for all integers $1 \leqslant t \leqslant n_0$,

$$\mathbb{P}\big\{U_{h,i}(t) \leqslant f^*\big\} \leqslant t\,(n_0)^{-4} \leqslant (n_0)^3\,.$$

Similarly, the same changes yield from Lemma 4.4 the following result for truncated HOO. For all integers $t \leqslant n_0$, for all suboptimal nodes (h,i) such that $h \leqslant D_{n_0}$ and $\Delta_{h,i} > \nu_1 \rho^h$, and for all integers $u \geqslant 1$ such that

$$u \geqslant \frac{8 \ln n_0}{(\Delta_{h,i} - \nu_1 \rho^h)^2}\,,$$

one has

$$\mathbb{P}\big\{U_{h,i}(t) > f^* \ \text{ and } \ T_{h,i}(t) > u\big\} \leqslant t\,(n_0)^{-4}\,.$$

Combining these three results (using the same methodology as in the proof of Lemma 4.5) shows that under Assumptions A1 and A2, for all suboptimal nodes (h,i) such that $h \leqslant D_{n_0}$ and $\Delta_{h,i} > \nu_1 \rho^h$, one has

$$\begin{aligned} \mathbb{E}[T_{h,i}(n_0)] &\leqslant \frac{8\,\ln n_0}{(\Delta_{h,i} - \nu_1 \rho^h)^2} + 1 + \sum_{t=u+1}^{n_0} \left(t\,(n_0)^4 + \sum_{s=1}^{\min\{D_{n_0}, n_0\}} (n_0)^{-3} \right) \\ &\leqslant \frac{8\,\ln n_0}{(\Delta_{h,i} - \nu_1 \rho^h)^2} + 3\,. \end{aligned}$$

(We thus even improve slightly the bound of Lemma 4.5.)

Adaptation of the proof of Theorem 4.1. The main change here comes from the fact that trees are cut at the depth D_{n_0}. As a consequence, the sets $\mathcal{I}_h$, $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{J}_h$ are defined only by referring to nodes of depth smaller than D_{n_0}. All steps of the proof can then be repeated, except the third step; there, while the bounds on the regret resulting from nodes of $\mathcal{T}^1$ and $\mathcal{T}^3$ go through without any changes (as these sets were constructed by considering all descendants of some base nodes), the bound on the regret $R_{n,2}$ associated with the nodes $\mathcal{T}^2$ calls for a modified proof since at this stage we used the property that each node is played at most once. But this is not true anymore for nodes (h,i) located at depth D_{n_0}, which can be played several times. Therefore the proof is modified as follows.

Consider a node at depth $h = D_{n_0}$. Then, by definition of D_{n_0},

$$h \geqslant D_{n_0} = \frac{(\ln n_0)/2 - \ln(1/\nu_1)}{\ln(1/\rho)}\,, \qquad \text{that is,} \qquad \nu_1\,\rho^h \leqslant \frac{1}{\sqrt{n_0}}\,.$$

Since the considered nodes are $2\nu_1\rho^{D_{n_0}}$–optimal, the corresponding domains are $4\nu_1\rho^{D_{n_0}}$–optimal by Lemma 4.1, thus also $4/\sqrt{n_0}$–optimal. The instantaneous regret incurred when playing any of these nodes is therefore bounded by $4/\sqrt{n_0}$; and the associated cumulative regret (over n_0 rounds)

can be bounded by $4\sqrt{n_0}$. In conclusion, with the notations of Theorem 4.1, we get the new bound

$$\mathbb{E}\big[R_{n,2}\big] \leqslant \sum_{h=0}^{H-1} 4\nu_1\rho^h \left|\mathcal{I}_h\right| + 4\sqrt{n_0} \leqslant 4\sqrt{n_0} + 4C\nu_1\nu_2^{-d'} \sum_{h=0}^{H-1} \rho^{h(1-d')} \,.$$

The rest of the proof goes through and only this additional additive factor of $4\sqrt{n_0}$ is suffered in the final regret bound. (The additional factor can be included in the O notation.)

6.3. Proof of Theorem 4.3 (regret bound for z–HOO). We start with the following equivalent of Lemma 4.1 in this new local context. Remember that h_0 is the smallest integer such that

$$2\nu_1\rho^{h_0} < \varepsilon_0 \,.$$

LEMMA 4.6. *Under Assumptions A1 and A2', for all $h \geqslant h_0$, if the suboptimality factor $\Delta_{h,i}$ of a region $\mathcal{P}_{h,i}$ is bounded by $c\nu_1\rho^h$ for some $c \in [0,2]$, then all arms in $\mathcal{P}_{h,i}$ are $L\max\{2c,\, c+1\}\,\nu_1\rho^h$–optimal, that is,*

$$\mathcal{P}_{h,i} \subset \mathcal{X}_{L\max\{2c,\, c+1\}\nu_1\rho^h} \,.$$

When $c = 0$, i.e., the node (h,i) is optimal, the bound improves to

$$\mathcal{P}_{h,i} \subset \mathcal{X}_{\nu_1\rho^h} \,.$$

PROOF. We first deal with the general case of $c \in [0,2]$. By the hypothesis on the suboptimality of $\mathcal{P}_{h,i}$, for all $\delta > 0$, there exists an element $x \in \mathcal{X}_{c\nu_1\rho^h+\delta} \cap \mathcal{P}_{h,i}$. If δ is small enough, e.g., $\delta \in \big(0,\, \varepsilon_0 - 2\nu_1\rho^{h_0}\big]$, then this element satisfies $x \in \mathcal{X}_{\varepsilon_0}$. Let $y \in \mathcal{P}_{h,i}$. By Assumption A1, $\ell(x,y) \leqslant \mathrm{diam}(\mathcal{P}_{h,i}) \leqslant \nu_1\rho^h$, which entails, by denoting $\varepsilon = \max\big\{0, \nu_1\rho^h - (f^* - f(x))\big\}$,

$$\ell(x,y) \leqslant \nu_1\rho^h \leqslant f^* - f(x) + \varepsilon \,, \qquad \text{that is,} \qquad y \in \mathcal{B}\big(x,\, f^* - f(x) + \varepsilon\big) \,.$$

Since $x \in \mathcal{X}_{\varepsilon_0}$ and $\varepsilon \leqslant \nu_1\rho^h \leqslant \nu_1\rho^{h_0} < \varepsilon_0$, the second part of Assumption A2' then yields

$$y \in \mathcal{B}\big(x,\, f^* - f(x) + \varepsilon\big) \;\subset \mathcal{X}_{L\left(2(f^*-f(x))+\varepsilon\right)} \,.$$

It follows from the definition of ε that $f^* - f(x) + \varepsilon = \max\big\{f^* - f(x),\, \nu_1\rho^h\big\}$, and this implies

$$y \in \mathcal{B}\big(x,\, f^* - f(x) + \varepsilon\big) \;\subset \mathcal{X}_{L\left(f^*-f(x)+\max\{f^*-f(x),\, \nu_1\rho^h\}\right)} \,.$$

But $x \in \mathcal{X}_{c\nu_1\rho^h+\delta}$, i.e., $f^* - f(x) \leqslant c\nu_1\rho^h + \delta$, we thus have proved

$$y \in \mathcal{X}_{L\left(\max\{2c,\, c+1\}\nu_1\rho^h+2\delta\right)} \,.$$

In conclusion, $\mathcal{P}_{h,i} \subset \mathcal{X}_{L\max\{2c,\, c+1\}\nu_1\rho^h+2L\delta}$ for all sufficiently small $\delta > 0$. Letting $\delta \to 0$ concludes the proof.

In the case of $c = 0$, we resort to the first part of Assumption A2', which can be applied since $\mathrm{diam}(\mathcal{P}_{h,i}) \leqslant \nu_1\rho^h \leqslant \varepsilon_0$ as already noted above, and can exactly be restated as indicating that for all $y \in \mathcal{P}_{h,i}$,

$$f^* - f(y) \leqslant \mathrm{diam}(\mathcal{P}_{h,i}) \leqslant \nu_1\rho^h \,;$$

that is, $\mathcal{P}_{h,i} \subset \mathcal{X}_{\nu_1\rho^h}$. □

We now provide an adaptation of Lemma 4.5 (actually based on adaptations of Lemmas 4.2 and 4.3), providing the same bound under slightly more restrictive conditions.

LEMMA 4.7. *Consider a depth $z \geqslant h_0$. Under Assumptions A1 and A2', the algorithm z–HOO satisfies that for all $n \geqslant 1$ and all suboptimal nodes (h,i) with $\Delta_{h,i} > \nu_1\rho^h$ and $h \geqslant z$,*

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant \frac{8\,\ln n}{(\Delta_{h,i} - \nu_1\rho^h)^2} + 4 \,.$$

PROOF. We consider some path (z, i_z^*), $(z+1, i_{z+1}^*)$, ... of optimal nodes, starting at depth z. We distinguish two cases, depending on whether there exists $z \leqslant k' \leqslant h-1$ such that $(h,i) \in \mathcal{C}(k', i_{k'}^*)$ or not.

In the first case, we denote k' the largest such k. The argument of Lemma 4.2 can be used without any change and shows that for all integers $u \geqslant 0$,

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant u + \sum_{t=u+1}^{n} \mathbb{P}\Big\{ \big[U_{s,i_s^*}(t) \leqslant f^* \text{ for some } s \in \{k+1, \dots, t-1\}\big] \\ \text{or} \quad \big[T_{h,i}(t) > u \ \text{ and } \ U_{h,i}(t) > f^*\big] \Big\}.$$

In the second case, we denote by (z, i_h) the ancestor of (h,i) located at depth z. By definition of z–HOO, $(H_t, I_t) \in \mathcal{C}(h,i)$ at some round $t \geqslant 1$ only if $B_{z,i_z^*}(t) \leqslant B_{z,i_h}(t)$ and since B–values can only increase on a chosen path, $(H_t, I_t) \in \mathcal{C}(h,i)$ can only happen if $B_{z,i_z^*}(t) \leqslant B_{h,i}(t)$. Repeating again the argument of Lemma 4.2, we get that for all integers $u \geqslant 0$,

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant u + \sum_{t=u+1}^{n} \mathbb{P}\Big\{ \big[U_{s,i_s^*}(t) \leqslant f^* \text{ for some } s \in \{z, \dots, t-1\}\big] \\ \text{or} \quad \big[T_{h,i}(t) > u \ \text{ and } \ U_{h,i}(t) > f^*\big] \Big\}.$$

Now, notice that Lemma 4.4 is valid without any assumption. On the other hand, with the modified assumptions, Lemma 4.3 is still true but only for optimal nodes (h,i) with $h \geqslant h_0$. Indeed, the only point in its proof where the assumptions were used was in the fourth line, when applying Lemma 4.1; here, Lemma 4.6 with $c = 0$ provides the needed guarantee.

The proof is concluded with the same computations as in the proof of Lemma 4.5. □

PROOF. **(of Theorem 4.3)** We simply follow the four steps in the proof of Theorem 4.1 with some slight adjustments. In particular, for $h \geqslant z$, we use the sets of nodes $\mathcal{I}_h$ and $\mathcal{J}_h$ defined therein.

First step. Lemma 4.7 bounds the expected number of times each node $(h,i) \in \mathcal{J}_h$ is visited. Since for these nodes $\Delta_{h,i} > 2\nu_1\rho^h$, we get

$$\mathbb{E}\big[T_{h,i}(n)\big] \leqslant \frac{8\,\ln n}{\nu_1^2 \rho^{2h}} + 4\,.$$

Second step. We bound here the cardinality $|\mathcal{I}_h|$. By Lemma 4.6 with $c = 2$, when $(h,i) \in \mathcal{I}_h$ and $h \geqslant z$, one has $\mathcal{P}_{h,i} \subset \mathcal{X}_{4L\nu_1\rho^h}$.

Now, by Assumption A1 and by using the same argument than in the second step of the proof of Theorem 4.1,

$$|\mathcal{I}_h| \leqslant \mathcal{N}\big(\mathcal{X}_{(4L\nu_1/\nu_2)\,\nu_2\rho^h},\, \ell,\, \nu_2\rho^h\big)\,.$$

Assumption A3 can be applied since $\nu_2\rho^h \leqslant 2\nu_1\rho^h \leqslant 2\nu_1\rho^{h_0} \leqslant \varepsilon_0$ and yields the inequality $|\mathcal{I}_h| \leqslant C\big(\nu_2\rho^h\big)^{-d}$.

Third step. We consider some integer $H \geqslant z$ to be defined by the analysis in the fourth step. We define a partition of the nodes located at a depth equal to or larger than z; more precisely,

- $\mathcal{T}^1$ contains the nodes of $\mathcal{I}_H$ and their descendants,
- $\mathcal{T}^2 = \bigcup_{z \leqslant h \leqslant H-1} \mathcal{I}_h$,
- $\mathcal{T}^3$ contains the nodes $\bigcup_{z+1 \leqslant h \leqslant H} \mathcal{J}_h$ and their descendants,

- $\mathcal{T}^4$ is formed by the nodes (z,i) located at depth z not belonging to $\mathcal{I}_z$, i.e., such that $\Delta_{z,i} > 2\nu_1\rho^z$, and their descendants.

As in the proof of Theorem 4.1 we denote by $R_{n,i}$ the regret resulting from the selection of nodes in $\mathcal{T}^i$, for $i \in \{1,2,3,4\}$.

Lemma 4.6 with $c = 2$ yields the bound $\mathbb{E}\big[R_{n,1}\big] \leqslant 4L\nu_1\rho^H n$, where we crudely bounded by n the number of times that nodes in $\mathcal{T}^1$ were played. Using that by definition each node of $\mathcal{T}^2$ can be played only once, we get

$$\mathbb{E}\big[R_{n,2}\big] \leqslant \sum_{h=z}^{H-1} \big(4L\nu_1\rho^h\big)\,|\mathcal{I}_h| \leqslant 4CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)}\,.$$

As for $R_{n,3}$, we also use here that nodes in $\mathcal{T}^3$ belong to some $\mathcal{J}_h$, with $z+1 \leqslant h \leqslant H$; in particular, they are the child of some element of $\mathcal{I}_{h-1}$ and as such, firstly, they are $4L\nu_1\rho^{h-1}$–optimal (by Lemma 4.6) and secondly, their number is bounded by $|\mathcal{J}_h| \leqslant 2|\mathcal{I}_{h-1}| \leqslant 2C\big(\nu_2\rho^{h-1}\big)^{-d}$. Thus,

$$\mathbb{E}\big[R_{n,3}\big] \leqslant \sum_{h=z+1}^{H} \big(4L\nu_1\rho^{h-1}\big) \sum_{i:(h,i)\in\mathcal{J}_h} \mathbb{E}\big[T_{h,i}(n)\big] \leqslant 8CL\nu_1\nu_2^{-d} \sum_{h=z+1}^{H} \rho^{(h-1)(1-d)} \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right),$$

where we used the bound of Lemma 4.7. Finally, for $\mathcal{T}^4$, we use that it contains at most $2^z - 1$ nodes, each of them being associated with a regret controlled by Lemma 4.7; therefore,

$$\mathbb{E}\big[R_{n,4}\big] \leqslant \big(2^z - 1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}} + 4\right).$$

Fourth step. Putting things together, we have proved that

$$\mathbb{E}\big[R_n\big] \leqslant 4L\nu_1\rho^H n + \mathbb{E}\big[R_{n,2}\big] + \mathbb{E}\big[R_{n,3}\big] + \big(2^z - 1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}} + 4\right),$$

where (using that $\rho < 1$ in the second inequality)

$$\begin{aligned}
&\mathbb{E}\big[R_{n,2}\big] + \mathbb{E}\big[R_{n,3}\big] \\
&\quad\leqslant\quad 4CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)} + 8CL\nu_1\nu_2^{-d} \sum_{h=z+1}^{H} \rho^{(h-1)(1-d)} \left(\frac{8\,\ln n}{\nu_1^2\rho^{2h}} + 4\right) \\
&\quad=\quad 4CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)} + 8CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)} \left(\frac{8\,\ln n}{\nu_1^2\rho^2\rho^{2h}} + 4\right) \\
&\quad\leqslant\quad 4CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)}\,\frac{1}{\rho^{2h}} + 8CL\nu_1\nu_2^{-d} \sum_{h=z}^{H-1} \rho^{h(1-d)} \left(\frac{8\,\ln n}{\nu_1^2\rho^2\rho^{2h}} + \frac{4}{\rho^{2h}}\right) \\
&\quad=\quad CL\nu_1\nu_2^{-d} \left(\sum_{h=z}^{H-1} \rho^{-h(1+d)}\right) \left(36 + \frac{64}{\nu_1^2\rho^2}\,\ln n\right).
\end{aligned}$$

Denoting

$$\gamma = \frac{4\,CL\nu_1\nu_2^{-d}}{(1/\rho)^{d+1} - 1} \left(\frac{16}{\nu_1^2\rho^2} + 9\right),$$

it follows that for $n \geqslant 2$

$$\mathbb{E}\big[R_{n,2}\big] + \mathbb{E}\big[R_{n,3}\big] \leqslant \gamma\,\rho^{-H(d+1)}\,\ln n\,.$$

It remains to define the parameter $H \geqslant z$. In particular, we propose to choose it such that the terms

$$4L\nu_1\rho^H n \qquad \text{and} \qquad \rho^{-H(d+1)}\,\ln n$$

are balanced. To this end, let H be the smallest integer k such that $4L\nu_1\rho^k n \leqslant \gamma\rho^{-k(d+1)}\ln n$; in particular,

$$\rho^H \leqslant \left(\frac{\gamma\ln n}{4L\nu_1 n}\right)^{1/(d+2)}$$

and

$$4L\nu_1\rho^{H-1}n > \gamma\rho^{-(H-1)(d+1)}\ln n\,, \qquad \text{implying} \qquad \gamma\,\rho^{-H(d+1)}\,\ln n \leqslant 4L\nu_1\rho^H n\;\rho^{-(d+2)}\,.$$

Note from the inequality that this H is such that

$$H \geqslant \frac{1}{d+2}\frac{\ln(4L\nu_1 n) - \ln(\gamma\ln n)}{\ln(1/\rho)}$$

and thus this H satisfies $H \geqslant z$ in view of the assumption of the theorem indicating that n is large enough. The final bound on the regret is then

$$\begin{aligned}
\mathbb{E}\big[R_n\big] &\leqslant 4L\nu_1\rho^H n + \gamma\,\rho^{-H(d+1)}\,\ln n + \big(2^z-1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}}+4\right)\\
&\leqslant \left(1+\frac{1}{\rho^{d+2}}\right)4L\nu_1\rho^H n + \big(2^z-1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}}+4\right)\\
&\leqslant \left(1+\frac{1}{\rho^{d+2}}\right)4L\nu_1 n\left(\frac{\gamma\ln n}{4L\nu_1 n}\right)^{1/(d+2)} + \big(2^z-1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}}+4\right)\\
&= \left(1+\frac{1}{\rho^{d+2}}\right)\big(4L\nu_1 n\big)^{(d+1)/(d+2)}(\gamma\ln n)^{1/(d+2)} + \big(2^z-1\big)\left(\frac{8\,\ln n}{\nu_1^2\rho^{2z}}+4\right).
\end{aligned}$$

This concludes the proof. □

6.4. Proof of Theorem 4.4 (regret bound for local HOO).

PROOF. We use the notation of the proof of Theorem 4.3. Let r_0 be a positive integer such that for $r \geqslant r_0$, one has

$$z_r \overset{\text{def}}{=} \lceil\log_2 r\rceil \geqslant h_0 \qquad \text{and} \qquad z_r \leqslant \frac{1}{d+2}\frac{\ln(4L\nu_1 2^r) - \ln(\gamma\ln 2^r)}{\ln(1/\rho)}\,;$$

we can therefore apply the result of Theorem 4.3 in regimes indexed by $r \geqslant r_0$. For previous regimes, we simply upper bound the regret by the number of rounds, that is, $2^{r_0}-2 \leqslant 2^{r_0}$. For round n, we denote by r_n the index of the regime where n lies in (regime $r_n = \lfloor\log_2(n+1)\rfloor$). Since regime r_n terminates at round $2^{r_n+1}-2$, we have

$$\begin{aligned}
\mathbb{E}\big[R_n\big] &\leqslant \mathbb{E}\big[R_{2^{r_n+1}-2}\big]\\
&\leqslant 2^{r_0} + \sum_{r=r_0}^{r_n}\left(1+\frac{1}{\rho^{d+2}}\right)\big(4L\nu_1 2^r\big)^{(d+1)/(d+2)}(\gamma\ln 2^r)^{1/(d+2)} + \big(2^{z_r}-1\big)\left(\frac{8\,\ln 2^r}{\nu_1^2\rho^{2z_r}}+4\right)\\
&\leqslant 2^{r_0} + C_1\,(\ln n)\sum_{r=r_0}^{r_n}\Big(2^{(d+1)/(d+2)}\Big)^r + \big(2/\rho^2\big)^{z_r}\\
&\leqslant 2^{r_0} + C_2\,(\ln n)\left(\Big(2^{(d+1)/(d+2)}\Big)^{r_n} + r_n\,\big(2/\rho^2\big)^{z_{r_n}}\right) \;= (\ln n)^2\;O\big(n^{(d+1)/(d+2)}\big)\,,
\end{aligned}$$

where $C_1, C_2 > 0$ denote some constants depending only on the parameters but not on n. Note that for the last equality we used that the first term in the sum of the two terms that depend on n dominates the second term. □

6.5. Proof of Theorem 4.5 (uniform upper bound on the regret of HOO against the class of all weak Lipschitz environments). Equations (4.4) and (4.5), which follow from Assumption A2, show that Assumption A2' is satisfied for $L = 2$ and all $\varepsilon_0 > 0$. We take, for instance, $\varepsilon_0 = 3\nu_1$. Moreover, since $\mathcal{X}$ has a packing dimension of D, all environments have a near-optimality dimension less than D. In particular, for all $D' > D$ (as shown in the second step of the proof of Theorem 4.1 in Section 6.1), there exists a constant C (depending only on ℓ, $\mathcal{X}$, $\varepsilon_0 = 3\nu_1$, ν_2, and D') such that Assumption A3 is satisfied. We can therefore take $h_0 = 0$ and apply Theorem 4.3 with $z = 0$ and $M \in \mathcal{F}_{\mathcal{X},\ell}$; the fact that all the quantities involved in the bound depend only on $\mathcal{X}$, ℓ, ν_2, D', and the parameters of HOO, but not on a particular environment in $\mathcal{F}$, concludes the proof.

6.6. Proof of Theorem 4.6 (minimax lower bound in metric spaces). Let $K \geqslant 2$ an integer to be defined later. We provide first an overview of the proof. Here, we exhibit a set $\mathcal{A}$ of environments for the $\{1, \dots, K+1\}$–armed bandit problem and a subset $\mathcal{F}' \subset \mathcal{F}_{\mathcal{X},\ell}$ which satisfy the following properties.

(i): The set $\mathcal{A}$ contains "difficult" environments for the $\{1, \dots, K+1\}$–armed bandit problem.

(ii): For any strategy $\varphi^{(\mathcal{X})}$ suited to the $\mathcal{X}$–armed bandit problem, one can construct a strategy $\psi^{(K+1)}$ for the $\{1, \dots, K+1\}$–armed bandit problem such that

$$\forall\ M \in \mathcal{F}',\ \exists\ \nu \in \mathcal{A}, \qquad \mathbb{E}_M\big[R_n(\varphi^{(\mathcal{X})})\big] = \mathbb{E}_\nu\big[R_n(\psi^{(K+1)})\big]\,.$$

We now provide the details.

PROOF. We only deal with the case of deterministic strategies. The extension to randomized strategies can be done using Fubini's theorem (by integrating also w.r.t. the auxiliary randomizations used).

First step. Let $\eta \in (0, 1/2)$ be a real number and $K \geqslant 2$ be an integer, both to be defined during the course of the analysis. The set $\mathcal{A}$ only contains K elements, denoted by $\nu^1, \dots, \nu^K$ and given by product distributions. For $1 \leqslant j \leqslant K$, the distribution ν^j is obtained as the product of the ν_i^j when $i \in \{1, \dots, K+1\}$ and where

$$\nu_i^j = \begin{cases} \mathrm{Ber}(1/2), & \text{if } i \neq j; \\ \mathrm{Ber}(1/2+\eta), & \text{if } i = j. \end{cases}$$

Rephrasing Lemma 2.2 in the context of this chapter we obtain:

LEMMA 4.8. *For all strategies $\psi^{(K+1)}$ for the $\{1, \dots, K+1\}$–armed bandit (where $K \geqslant 2$), one has*

$$\max_{j=1,\dots,K} \mathbb{E}_{\nu^j}\big[R_n(\psi^{(K+1)})\big] \geqslant n\eta\left(1 - \frac{1}{K} - \eta\sqrt{4\ln(4/3)}\sqrt{\frac{n}{K}}\right).$$

Second step. We now need to construct $\mathcal{F}'$ such that item (ii) is satisfied. We assume that K is such that $\mathcal{X}$ contains K disjoint balls with radius η. (We shall quantify later in this proof a suitable value of K.) Denoting by $x_1, \dots, x_K$ the corresponding centers, these disjoint balls are then $\mathcal{B}(x_1, \eta)$, …, $\mathcal{B}(x_K, \eta)$.

With each of these balls we now associate a bandit environment over $\mathcal{X}$, in the following way. For all $x^* \in \mathcal{X}$, we introduce a mapping $g_{x^*,\eta}$ on $\mathcal{X}$ defined by

$$g_{x^*,\eta}(x) = \max\bigl\{0,\ \eta - \ell(x,x^*)\bigr\}$$

for all $x \in \mathcal{X}$. This mapping is used to define an environment $M_{x^*,\eta}$ over $\mathcal{X}$, as follows. For all $x \in \mathcal{X}$,

$$M_{x^*,\eta}(x) = \mathrm{Ber}\left(\frac{1}{2} + g_{x^*,\eta}(x)\right).$$

Let $f_{x^*,\eta}$ be the corresponding mean-payoff function; its values equal

$$f_{x^*,\eta}(x) = \frac{1}{2} + \max\bigl\{0,\ \eta - \ell(x,x^*)\bigr\}$$

for all $x \in \mathcal{X}$. Note that the mean payoff is maximized at $x = x^*$ (with value $1/2 + \eta$) and is minimal for all points lying outside $\mathcal{B}(x^*,\eta)$, with value $1/2$. In addition, that ℓ is a metric entails that these mean-payoff functions are 1–Lipschitz and thus are also weakly Lipschitz. (This is the only point in the proof where we use that ℓ is a metric.) In conclusion, we consider

$$\mathcal{F}' = \bigl\{M_{x_1,\eta},\ \ldots,\ M_{x_K,\eta}\bigr\} \ \subset \mathcal{F}_{\mathcal{X},\ell}\,.$$

Third step. We describe how to associate with each (deterministic) strategy $\varphi^{(\mathcal{X})}$ on $\mathcal{X}$ a (random) strategy $\psi^{(K+1)}$ on the finite set of arms $\{1,\ldots,K+1\}$. Each of these strategies is indeed given by a sequence of mappings,

$$\varphi_1^{(\mathcal{X})}, \varphi_2^{(\mathcal{X})}, \ldots \qquad \text{and} \qquad \psi_1^{(K+1)}, \psi_2^{(K+1)}, \ldots$$

where for $t \geqslant 1$, the mappings $\varphi_t^{(\mathcal{X})}$ and $\psi_t^{(K+1)}$ should only depend on the past up to the beginning of round t. Since the strategy $\varphi^{(\mathcal{X})}$ is deterministic, the mapping $\varphi_t^{(\mathcal{X})}$ takes only into account the past rewards $Y_1,\ldots,Y_{t-1}$ and is therefore a mapping $[0,1]^{t-1} \to \mathcal{X}$. (In particular, $\varphi_1^{(\mathcal{X})}$ equals a constant.)

We use the notations I'_t and Y'_t for, respectively, the arms pulled and the rewards obtained by the strategy $\psi^{(K+1)}$ at each round t. The arms I'_t are drawn at random according to the distributions

$$\psi_t^{(K+1)}\bigl(I'_1,\ldots,I'_{t-1},\,Y'_1,\ldots,Y'_{t-1}\bigr)\,,$$

which we now define. (Actually, they will depend on the obtained payoffs $Y'_1,\ldots,Y'_{t-1}$ only.) To do that, we need yet another mapping T that links elements in $\mathcal{X}$ to probability distributions over $\{1,\ldots,K+1\}$. Denoting by δ_k the Dirac probability on $k \in \{1,\ldots,K+1\}$, the mapping T is defined as

$$T(x) = \begin{cases} \delta_{K+1}\,, & \text{if } x \notin \displaystyle\bigcup_{j=1,\ldots,K} \mathcal{B}(x_j,\eta); \\ \left(1 - \dfrac{\ell(x,x_j)}{\eta}\right)\delta_j + \dfrac{\ell(x,x_j)}{\eta}\,\delta_{K+1}\,, & \text{if } x \in \mathcal{B}(x_j,\eta) \text{ for some } j \in \{1,\ldots,K\}, \end{cases}$$

for all $x \in \mathcal{X}$. Note that this definition is legitimate because the balls $\mathcal{B}(x_j,\eta)$ are disjoint when j varies between 1 and K.

Finally, $\psi^{(K+1)}$ is defined as follows. For all $t \geqslant 1$,

$$\psi_t^{(K+1)}\bigl(I'_1,\ldots,I'_{t-1},\,Y'_1,\ldots,Y'_{t-1}\bigr) = \psi_t^{(K+1)}\bigl(Y'_1,\ldots,Y'_{t-1}\bigr) = T\Bigl(\varphi_t^{(\mathcal{X})}\bigl(Y'_1,\ldots,Y'_{t-1}\bigr)\Bigr)\,.$$

Before we proceed, we study the distribution of the reward Y' obtained under ν^i (for $i \in \{1,\ldots,K\}$) by the choice of a random arm I' drawn according to $T(x)$, for some $x \in \mathcal{X}$. Since Y' can only take the values 0 or 1, its distribution is a Bernoulli distribution whose parameter $\mu_i(x)$ we compute now. The computation is based on the fact that under ν^i, the Bernoulli distribution

corresponding to arm j has $1/2$ as an expectation, except if $j = i$, in which case it is $1/2 + \eta$. Thus, for all $x \in \mathcal{X}$,

$$\mu_i(x) = \begin{cases} 1/2\,, & \text{if } x \notin \mathcal{B}(x_i, \eta); \\ \left(1 - \dfrac{\ell(x, x_i)}{\eta}\right)\left(\dfrac{1}{2} + \eta\right) + \dfrac{\ell(x, x_i)}{\eta}\,\dfrac{1}{2} = \dfrac{1}{2} + \eta - \ell(x, x_i)\,, & \text{if } x \in \mathcal{B}(x_i, \eta). \end{cases}$$

That is, $\mu_i = f_{x_i,\eta}$ on $\mathcal{X}$.

Fourth step. We now prove that the distributions of the regrets of $\varphi^{(\mathcal{X})}$ under $M_{x_j,\eta}$ and of $\psi^{(K+1)}$ under ν^j are equal for all $j = 1, \dots, K$. On the one hand, the expectations of rewards associated with the best arms equal $1/2 + \eta$ under the two environments. On the other hand, one can prove by induction that the sequences $Y_1, Y_2, \dots$ and $Y'_1, Y'_2, \dots$ have the same distribution. (In the argument below, conditioning by empty sequences means no conditioning. This will be the case only for $t = 1$.)

For all $t \geqslant 1$, we denote

$$X'_t = \varphi_t^{(\mathcal{X})}\big(Y'_1, \dots, Y'_{t-1}\big)\,.$$

Under ν^j and given $Y'_1, \dots, Y'_{t-1}$, the distribution of Y'_t is obtained by definition as the two-step random draw of $I'_t \sim T(X'_t)$ and then, conditionally on this first draw, $Y'_t \sim \nu^j_{I'_t}$. By the above results, the distribution of Y'_t is thus a Bernoulli distribution with parameter $\mu_j(X'_t)$.

At the same time, under $M_{x_j,\eta}$ and given $Y_1, \dots, Y_{t-1}$, the choice of

$$X_t = \varphi_t^{(\mathcal{X})}\big(Y_1, \dots, Y_{t-1}\big)$$

yields a reward Y_t distributed according to $M_{x_j,\eta}(X_t)$, that is, by definition and with the notations above, a Bernoulli distribution with parameter $f_{x_j,\eta}(X_t) = \mu_j(X_t)$.

The argument is concluded by induction and by using the fact that rewards are drawn independently in each round.

Fifth step. We summarize what we proved so far. For $\eta \in (0, 1/2)$, provided that there exist $K \geqslant 2$ disjoint balls $\mathcal{B}(x_j, \eta)$ in $\mathcal{X}$, we could construct, for all strategies $\varphi^{(\mathcal{X})}$ for the $\mathcal{X}$–armed bandit problem, a strategy $\psi^{(K+1)}$ for the $\{1, \dots, K+1\}$–armed bandit problem such that, for all $j = 1, \dots, K$ and all $n \geqslant 1$,

$$\mathbb{E}_{M_{x_j,\eta}}\big[R_n(\varphi^{(\mathcal{X})})\big] = \mathbb{E}_{\nu^j}\big[R_n(\psi^{(K+1)})\big]\,.$$

But by the assumption on the packing dimension, there exists $c > 0$ such that for all $\eta < 1/2$, the choice of $K_\eta = \lceil c\,\eta^{-D} \rceil \geqslant 2$ guarantees the existence of such K_η disjoint balls. Substituting this value, and using the results of the first and fourth steps of the proof, we get

$$\max_{j=1,\dots,K_\eta} \mathbb{E}_{M_{x_j,\eta}}\big[R_n(\varphi^{(\mathcal{X})})\big] = \max_{j=1,\dots,K_\eta} \mathbb{E}_{\nu^j}\big[R_n(\psi^{(K+1)})\big] \geqslant n\eta\left(1 - \frac{1}{K_\eta} - \eta\sqrt{4\ln(4/3)}\sqrt{\frac{n}{K_\eta}}\right)\,.$$

The proof is concluded by noting that

- the left-hand side is smaller than the maximal regret w.r.t. all weak-Lipschitz environments;
- the right-hand side can be lower bounded and then optimized over $\eta < 1/2$ in the following way.

By definition of K_η and the fact that it is larger than 2, one has

$$n\eta\left(1 - \frac{1}{K_\eta} - \eta\sqrt{4\ln(4/3)}\sqrt{\frac{n}{K_\eta}}\right)$$

$$\geqslant n\eta\left(1-\frac{1}{2}-\eta\sqrt{4\ln(4/3)}\sqrt{\frac{n}{c\eta^{-D}}}\right)=n\eta\left(\frac{1}{2}-C\,\eta^{1+D/2}\sqrt{n}\right)$$

where $C=\sqrt{\big(4\ln(4/3)\big)\,/\,c}$. We can optimize the final lower bound over $\eta\in[0,\,1/2]$.

To that end, we choose, for instance, η such that $C\,\eta^{1+D/2}\sqrt{n}=1/4$, that is,

$$\eta=\left(\frac{1}{4C\sqrt{n}}\right)^{1/(1+D/2)}=\left(\frac{1}{4C}\right)^{1/(1+D/2)}n^{-1/(D+2)}\,.$$

This gives the lower bound

$$\frac{1}{4}\left(\frac{1}{4C}\right)^{1/(1+D/2)}n^{1-1/(D+2)}=\underbrace{\frac{1}{4}\left(\frac{1}{4C}\right)^{1/(1+D/2)}}_{=\,\gamma(c,D)}n^{(D+1)/(D+2)}\,.$$

To ensure that this choice of η is valid we need to show that $\eta\leqslant 1/2$. Since the latter requirement is equivalent to

$$n\geqslant\left(2\left(\frac{1}{4C}\right)^{1/(1+D/2)}\right)^{D+2}\,,$$

it suffices to choose the right-hand side to be $N(c,D)$; we then get that $\eta\leqslant 1/2$ indeed holds for all $n\geqslant N(c,D)$, thus concluding the proof of the theorem. □

CHAPTER 5

Open-Loop Optimistic Planning

We consider the problem of planning in a stochastic and discounted environment with a limited numerical budget. More precisely, we investigate strategies exploring the set of possible sequences of actions, so that, once all available numerical resources (e.g. CPU time, number of calls to a generative model) have been used, one returns a recommendation on the best possible immediate action (or sequence of actions) to follow based on this exploration. The performance of a strategy is assessed in terms of its simple regret, that is the loss in performance resulting from choosing the recommended action instead of an optimal one. We first provide a minimax lower bound for this problem, and show that a uniform planning strategy matches this minimax rate (up to a logarithmic factor). Then we propose a UCB (Upper Confidence Bounds)-based planning algorithm, called OLOP (Open-Loop Optimistic Planning), which is also minimax optimal, and prove that it enjoys much faster rates when there is a small proportion of near-optimal sequences of actions. Finally, we compare our results with the regret bounds one can derive for our setting with bandits algorithms designed for an infinite number of arms.

Contents

This chapter is a joint work with Rémi Munos. It is based on the paper Bubeck and Munos [2010] published in the proceedings of the 23rd Annual Conference on Learning Theory.

1. Introduction

We consider the problem of planning in general stochastic and discounted environments. More precisely, the decision making problem consists in an exploration phase followed by a recommendation. First, the agent explores freely the set of possible sequences of actions, using a finite budget of n actions. Then the agent makes a recommendation on the first action $a(n)$ (or sequence of actions) to play. This decision making problem is described precisely in Figure 1. The goal of the agent is to find the best way to explore its environment (first phase) so that, once the available

Exploration in a stochastic and discounted environment.

Parameters available to the agent: discount factor $\gamma \in (0,1)$, number of actions K, number of rounds n.

Parameters unknown to the agent: the reward distributions $\nu(a)$, $a \in A^*$.

For each episode $m \geqslant 1$; for each moment in the episode $t \geqslant 1$;

(1) If n actions have already been performed then the agent outputs an action (or a sequence) $a(n)$ and the game stops.
(2) The agent chooses an action $a_t^m \in A$.
(3) The environment draws $Y_t^m \sim \nu(a_{1:t}^m)$ and the agent receives the reward $\gamma^t Y_t^m$.
(4) The agent decides to either move the next moment $t+1$ in the episode or to reset to its initial position and move the next episode $m+1$.

Goal: maximize the value of the recommended action (or sequence): $V(a(n))$ (see (5.1) for the definition of the value of an action).

Figure 1: Exploration in a stochastic and discounted environment.

resources have been used, he is able to make the best possible recommendation on the action to play in the environment.

During the exploration of the environment, the agent iteratively selects sequences of actions and receives a reward after each action. More precisely, at time step t during the m^{th} sequence, the agent played $a_{1:t}^m = a_1^m \dots a_t^m \in A^t$ and receives a discounted reward $\gamma^t Y_t^m$ where $\gamma \in (0,1)$ is the discount factor. We make a stochastic assumption on the generating process for the reward: given $a_{1:t}^m$, Y_t is drawn from a probability distribution $\nu(a_{1:t}^m)$ on $[0,1]$. Given $a \in A^t$, we write $\mu(a)$ for the mean of the probability $\nu(a)$.

The performance of the recommended action $a(n) \in A$ (or sequence, in which case $a(n) \in A^h$) is assessed in terms of the so-called **simple regret** r_n, which is the performance loss resulting from choosing this sequence and then following an optimal path instead of following an optimal path from the beginning:

$$r_n = V - V(a(n)),$$

where $V(a(n))$ is the (discounted) value of the action (or sequence) $a(n)$, defined for any finite sequence of actions $a \in A^h$ as:

$$V(a) = \sup_{u \in A^\infty : u_{1:h} = a} \sum_{t \geqslant 1} \gamma^t \mu(u_{1:t}), \tag{5.1}$$

and V is the optimal value, that is the maximum expected sum of discounted rewards one may obtain (i.e. the sup in (5.1) is taken over all sequences in A^∞).

Note that this simple regret criterion will be extensively studied in the context of multi-armed bandit in Chapter 6 and Chapter 7.

An important application of this framework concerns the problem of planning in Markov Decision Processes (MDPs) with very large state spaces. We assume that the agent possesses a generative model which enables to generate a reward and a transition from any state-action to a next state, according to the underlying reward and transition model of the MDP. In this context, we propose to use the generative model to perform a planning from the current state (using a finite

budget of n calls to a generative model) to generate a near-optimal action $a(n)$ and then apply $a(n)$ in the real environment. This action modifies the environment and the planning procedure is repeated from the next state to select the next action and so on. From each state, the planning consists in the exploration of the set of possible sequences of actions as described in Figure 1, where the generative model is used to generate the rewards.

Note that, using control terminology, the setting described above (from a given state) is called "open-loop" planning, because the class of considered policies (i.e. sequences of actions) are only function of time (and not of the underlying resulting states). This open-loop planning is in general sub-optimal compared to the optimal (closed-loop) policy (mapping from states to actions). However, here, while the planning is open-loop (i.e. we do not take into consideration the subsequent states in the planning), the resulting policy is closed-loop (since the chosen action depends on the current state).

This approach to MDPs has already been investigated as an alternative to usual dynamic programming approaches (which approximate the optimal value function to design a near optimal policy) to circumvent the computational complexity issues. For example, Kearns et al. describe a sparse sampling method that uses a finite amount of computational resources to build a look-ahead tree from the current state, and returns a near-optimal action with high probability.

Another field of application is POMDPs (Partially Observable Markov Decision Problems), where from the current belief state an open-loop plan may be built to select a near-optimal immediate action (see e.g. Yu et al. [2005], Hsu et al. [2007]). Note that, in these problems, it is very common to have a limited budget of computational resources (CPU time, memory, number of calls to the generative model, ...) to select the action to perform in the real environment, and we aim at making an efficient use of the available resources to perform the open-loop planning.

Moreover, in many situations, the generation of state-transitions is computationally expensive, thus it is critical to make the best possible use of the available number of calls to the model to output the action. For instance, an important problem in waste-water treatment concerns the control of a biochemical process for anaerobic digestion. The chemical reactions involve hundreds of different bacteria and the simplest models of the dynamics already involve dozens of variables (for example, the well-known model called ADM1 Batstone et al. [2002] contains 32 state variables) and their simulation is numerically heavy. Because of the curse of dimensionality, it is impossible to compute an optimal policy for such model. The methodology described above aims at a less ambitious goal, and search for a closed-loop policy which is open-loop optimal at each time step. While this policy is suboptimal, it is also a more reasonable target in terms of computational complexity. The strategy considered here proposes to use the model to simulate transitions and perform a complete open-loop planning at each time step.

The main contribution of this chapter is the analysis of an adaptive exploration strategy of the search space, called Open-Loop Optimistic Planning (OLOP), which is based on the "optimism in the face of uncertainty" principle, i.e. where the most promising sequences of actions are explored first. The idea of optimistic planning has already been investigated in the simple case of deterministic environments, Hren and Munos [2008]. Here we consider the non-trivial extension of this optimistic approach to planning in stochastic environments. For that purpose, upper confidence bounds (UCBs) are assigned to all sequences of actions, and the exploration expands further the sequences with highest UCB. The idea of selecting actions based on UCBs comes from the multi-armed bandits literature, see Lai and Robbins [1985], Auer et al. [2002]. Planning under uncertainty using UCBs has been considered previously in Chang et al. [2007] (the so-called UCB sampling) and in Kocsis and Szepesvari [2006], where the resulting algorithm, UCT (UCB applied

to Trees), has been successfully applied to the large scale tree search problem of computer-go, see Gelly et al. [2006]. However, its regret analysis shows that UCT may perform very poorly because of overly-optimistic assumptions in the design of the bounds, see Coquelin and Munos [2007]. This work is close in spirit to BAST (Bandit Algorithm for Smooth Trees), Coquelin and Munos [2007], the Zooming Algorithm, Kleinberg et al. [2008a] and the strategy HOO described in Chapter 4. Like in these previous works, the performance bounds of OLOP are expressed in terms of a measure of the proportion of near-optimal paths.

However, as we shall discuss in Section 4, these previous algorithms fail to obtain minimax guarantees for our problem. Indeed, a particularity of our planning problem is that the value of a sequence of action is defined as the sum of discounted rewards along the path, thus the rewards obtained along any sequence provides information, not only about that specific sequence, but also about any other sequence sharing the same initial actions. OLOP is designed to use this property as efficiently as possible, to derive tight upper-bounds on the value of each sequence of actions.

Note that our results does not compare with traditional regret bounds for MDPs, such as the ones proposed in Auer et al. [2009]. Indeed, in this case one compares to the optimal closed-loop policy, and the resulting regret usually depends on the size of the state space (as well as on other parameters of the MDP).

Outline. We exhibit in Section 2 the minimax rate (up to a logarithmic factor) for the simple regret in discounted and stochastic environments: both lower and upper bounds are provided. Then in Section 3 we describe the OLOP strategy, and show that if there is a small proportion of near-optimal sequences of actions, then faster rates than minimax can be derived. In Section 4 we compare our results with previous works and present several open questions. Finally Section 5 gathers the proofs.

Notations To shorten the equations we use several standard notations over alphabets. We collect them here: $A^0 = \{\emptyset\}$, A^* is the set of finite words over A (including the null word $\emptyset$), for $a \in A^*$ we note $h(a)$ the integer such that $a \in A^{h(a)}$, $aA^h = \{ab, b \in A^h\}$, for $a \in A^h$ and $h' > h$ we note $a_{1:h'} = a\emptyset \dots \emptyset$ and $a_{1:0} = \emptyset$.

2. Minimax optimality

In this section we derive a lower bound on the simple regret (in the worst case) of any agent, and propose a simple (uniform) forecaster which attains this optimal minimax rate (up to a logarithmic factor). The main purpose of the section on the uniform planning is to show explicitly the special concentrations property that our model enjoys.

2.1. Minimax lower bound. We propose here a new lower bound, whose proof can be found in Appendix 5.1 and which is based on the technique developed in Auer et al. [2003]. Note that this lower bound is not a particular case of the ones derived in Kleinberg et al. [2008a] or Bubeck et al. [2009c] in a more general framework, as we shall see in Section 4.

THEOREM 5.1. *Any agent satisfies:*

$$\sup_{\nu} \mathbb{E} r_n = \begin{cases} \Omega\left(\left(\frac{\log n}{n} \right)^{\frac{\log 1/\gamma}{\log K}} \right) & \text{if } \gamma\sqrt{K} > 1, \\ \Omega\left(\sqrt{\frac{\log n}{n}} \right) & \text{if } \gamma\sqrt{K} \leqslant 1. \end{cases}$$

2.2. Uniform Planning. To start gently, let us consider first (and informally) a *naive version* of the uniform planning. One can choose a depth H, uniformly test all sequences of actions in A^H (with $(n/H)/K^H$ samples for each sequence), and then return the empirical best sequence. Cutting the sequences at depth H implies an error of order γ^H, and relying on empirical estimates with $(n/H)/K^H$ samples adds an error of order $\sqrt{\frac{HK^H}{n}}$, leading to a simple regret bounded as $O\left(\gamma^H + \sqrt{\frac{HK^H}{n}}\right)$. Optimizing over H yields an upper bound on the simple regret of the naive uniform planning of order:

$$O\left(\left(\frac{\log n}{n}\right)^{\frac{\log 1/\gamma}{\log K + 2\log 1/\gamma}}\right), \tag{5.2}$$

which does not match the lower bound. The cautious reader probably understands why this version of uniform planning is suboptimal. Indeed we do not use the fact that any sequence of actions of the form ab gives information on the sequences ac. Hence, the concentration of the empirical mean for short sequences of actions is much faster than for long sequences. This is the critical property which enables us to fasten the rates with respect to traditional methods, see Section 4 for more discussion on this.

We describe now the *good version* of uniform planning. Let $H \in \mathbb{N}$ be the largest integer such that $HK^H \leqslant n$. Then the procedure goes as follows: For each sequence of actions $a \in A^H$, the uniform planning allocates one episode (of length H) to estimate the value of the sequence a, that is it receives $Y_t^a \sim \nu(a_{1:t})$, $1 \leqslant t \leqslant H$ (drawn independently). At the end of the allocation procedure, it computes for all $a \in A^h$, $h \leqslant H$, the empirical average reward of the sequence a:

$$\widehat{\mu}(a) = \frac{1}{K^{H-h}} \sum_{b \in A^H : b_{1:h} = a} Y_h^b.$$

(obtained with K^{H-h} samples.) Then, for all $a \in A^H$, it computes the empirical value of the sequence a:

$$\widehat{V}(a) = \sum_{t=1}^{H} \gamma^t \widehat{\mu}(a_{1:t}).$$

It outputs $a(n) \in A$ defined as the first action of the sequence $\arg\max_{a \in A^H} \widehat{V}(a)$ (ties break arbitrarily).

This version of uniform planning makes a much better use of the reward samples than the naive version. Indeed, for any sequence $a \in A^h$, it collects the rewards Y_h^b received for sequences $b \in aA^{H-h}$ to estimate $\mu(a)$. Since $|aA^{H-h}| = K^{H-h}$, we obtain an estimation error for $\mu(a)$ of order $\sqrt{K^{h-H}}$. Then, thanks to the discounting, the estimation error for $V(a)$, with $a \in A^H$, is of order $K^{-H/2} \sum_{h=1}^{H} (\gamma\sqrt{K})^h$. On the other hand, the approximation error for cutting the sequences at depth H is still of order γ^H. Thus, since H is the largest depth (given n and K) at which we can explore once each node, we obtain the following behavior: When K is large, precisely $\gamma\sqrt{K} > 1$, then H is small and the estimation error is of order γ^H, resulting in a simple regret of order $n^{-(\log 1/\gamma)/\log K}$. On the other hand, if γ is small, precisely $\gamma\sqrt{K} < 1$, then the depth H becomes less important, and the estimation error is of order $K^{-H/2}$, resulting in a simple regret of order $n^{-1/2}$. This reasoning is made precise in Appendix 5.2 (supplementary material section), where we prove the following Theorem.

THEOREM 5.2. *The (good) uniform planning satisfies:*

$$\mathbb{E}r_n \leqslant \begin{cases} O\left(\sqrt{\log n}\left(\frac{\log n}{n}\right)^{\frac{\log 1/\gamma}{\log K}}\right) & \text{if } \gamma\sqrt{K} > 1, \\ O\left(\frac{(\log n)^2}{\sqrt{n}}\right) & \text{if } \gamma\sqrt{K} = 1, \\ O\left(\frac{\log n}{\sqrt{n}}\right) & \text{if } \gamma\sqrt{K} < 1. \end{cases}$$

REMARK 5.1. *We do not know whether the $\sqrt{\log n}$ (respectively $(\log n)^{3/2}$ in the case $\gamma\sqrt{K} = 1$) gap between the upper and lower bound comes from a suboptimal analysis (either in the upper or lower bound) or from a suboptimal behavior of the uniform forecaster.*

3. OLOP (Open Loop Optimistic Planning)

The uniform planning described in Section 2.2 is a static strategy, it does not adapt to the rewards received in order to improve its exploration. A stronger strategy could select, at each round, the next sequence to explore as a function of the previously observed rewards. In particular, since the value of a sequence is the sum of discounted rewards, one would like to explore more intensively the sequences starting with actions that already yielded high rewards. In this section we describe an adaptive exploration strategy, called Open Loop Optimistic Planning (OLOP), which explores first the most promising sequences, resulting in much stronger guarantees than the one derived for uniform planning.

OLOP proceeds as follows. It assigns upper confidence bounds (UCBs), called B-values, to all sequences of actions, and selects at each round a sequence with highest B-value. This idea of a UCB-based exploration comes from the multi-armed bandits literature, see Auer et al. [2002]. It has already been extended to hierarchical bandits, Chang et al. [2007], Kocsis and Szepesvari [2006], Coquelin and Munos [2007], and to bandits in metric (or even more general) spaces, Auer et al. [2007], Kleinberg et al. [2008a], Bubeck et al. [2009c].

Like in these previous works, we express the performance of OLOP in terms of a measure of the proportion of near-optimal paths. More precisely, we define $\kappa_c \in [1, K]$ as the branching factor of the set of sequences in A^h that are $c - \frac{\gamma^{h+1}}{1-\gamma}$-optimal, where c is a positive constant, i.e.

$$\kappa_c = \limsup_{h\to\infty} \left| \left\{ a \in A^h : V(a) \geqslant V - c\frac{\gamma^{h+1}}{1-\gamma} \right\} \right|^{1/h}. \tag{5.3}$$

Intuitively, the set of sequences $a \in A^h$ that are $\frac{\gamma^{h+1}}{1-\gamma}$-optimal are the sequences for which the perfect knowledge of the discounted sum of mean rewards $\sum_{t=1}^h \gamma^t \mu(a_{1:t})$ is not sufficient to decide whether a belongs to an optimal path or not, because of the unknown future rewards for $t > h$. In the main result, we consider κ_2 (rather than κ_1) to account for an additional uncertainty due to the empirical estimation of $\sum_{t=1}^h \gamma^t \mu(a_{1:t})$. In Section 4, we discuss the link between κ and the other measures of the set of near-optimal states introduced in the previously mentioned works.

3.1. The OLOP algorithm. The OLOP algorithm is described in Figure 2. It makes use of some B-values assigned to any sequence of actions in A^L. At time $m = 0$, the B-values are initialized to $+\infty$. Then, after episode $m \geqslant 1$, the B-values are defined as follows: For any $1 \leqslant h \leqslant L$, for any $a \in A^h$, let

$$T_a(m) = \sum_{s=1}^{m} \mathbb{1}\{a_{1:h}^s = a\}$$

Open Loop Optimistic Planning:

Let M be the largest integer such that $M\lceil \log M/(2\log 1/\gamma)\rceil \leqslant n$. Let $L = \lceil \log M/(2\log 1/\gamma)\rceil$.

For each episode $m = 1, 2, \ldots, M$;

(1) The agent computes the B–values at time $m-1$ for sequences of actions in A^L (see Section 3.1) and chooses a sequence that maximizes the corresponding B–value:

$$a^m \in \operatorname*{argmax}_{a\in A^L} B_a(m-1).$$

(2) The environment draws the sequence of rewards $Y_t^m \sim \nu(a_{1:t}^m)$, $t = 1, \ldots, L$.

Return an action that has been the most played: $a(n) = \operatorname{argmax}_{a\in A} T_a(M)$.

Figure 2: Open Loop Optimistic Planning

be the number of times we played a sequence of actions beginning with a. Now we define the empirical average of the rewards for the sequence a as:

$$\widehat{\mu}_a(m) = \frac{1}{T_a(m)} \sum_{s=1}^{m} Y_h^s \mathbb{1}\{a_{1:h}^s = a\},$$

if $T_a(m) > 0$, and 0 otherwise. The corresponding upper confidence bound on the value of the sequence of actions a is by definition:

$$U_a(m) = \sum_{t=1}^{h} \gamma^t \widehat{\mu}_{a_{1:t}}(m) + \gamma^t \sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} + \frac{\gamma^{h+1}}{1-\gamma},$$

if $T_a(m) > 0$ and $+\infty$ otherwise. Now that we have upper confidence bounds on the value of many sequences of actions we can sharpen these bounds for the sequences $a \in A^L$ by defining the B-values as:

$$B_a(m) = \inf_{1\leqslant h\leqslant L} U_{a_{1:h}}(m).$$

At each episode $m = 1, 2, \ldots, M$, OLOP selects a sequence $a^m \in A^L$ with highest B–value, observes the rewards $Y_t^m \sim \nu(a_{1:t}^m)$, $t = 1, \ldots, L$ provided by the environment, and updates the B–values. At the end of the exploration phase, OLOP returns an action that has been the most played, *i.e.* $a(n) = \operatorname{argmax}_{a\in A} T_a(M)$.

3.2. Main result.

THEOREM 5.3 (Main Result). *Let $\kappa_2 \in [1, K]$ be defined by (5.3). Then, for any $\kappa' > \kappa_2$, OLOP satisfies:*

$$\mathbb{E}r_n = \begin{cases} \widetilde{O}\left(n^{-\frac{\log 1/\gamma}{\log \kappa'}}\right) & \text{if } \gamma\sqrt{\kappa'} > 1, \\ \widetilde{O}\left(n^{-\frac{1}{2}}\right) & \text{if } \gamma\sqrt{\kappa'} \leqslant 1. \end{cases}$$

(We say that $u_n = \widetilde{O}(v_n)$ if there exists $\alpha, \beta > 0$ such that $u_n \leqslant \alpha(\log(v_n))^\beta v_n$)

REMARK 5.2. *One can see that the rate proposed for OLOP greatly improves over the uniform planning whenever there is a small proportion of near-optimal paths (i.e. κ is small). Note that this does not contradict the lower bound proposed in Theorem 5.1. Indeed κ provides a description*

of the environment ν, and the bounds are expressed in terms of that measure, one says that the bounds are distribution-dependent. Nonetheless, OLOP does not require the knowledge of κ, thus one can take the supremum over all $\kappa \in [1, K]$, and see that it simply replaces κ by K, proving that OLOP is minimax optimal (up to a logarithmic factor).

REMARK 5.3. *In the analysis of OLOP, we relate the simple regret to the more traditional* **cumulative regret**, *defined at round n as $R_n = \sum_{m=1}^{M} \left(V - V(a^m) \right)$. Indeed, in the proof of Theorem 5.3, we first show that $r_n = \widetilde{O}\left(\frac{R_n}{n}\right)$, and then we bound (in expectation) this last term. Thus the same bounds apply to $\mathbb{E}R_n$ with a multiplicative factor of order n. In this chaper, we focus on the simple regret rather than on the traditional cumulative regret because we believe that it is a more natural performance criterion for the planning problem considered here. However note that OLOP is also minimax optimal (up to a logarithmic factor) for the cumulative regret, since one can also derive lower bounds for this performance criterion using the proof of Theorem 5.1.*

REMARK 5.4. *One can also see that the analysis carries over to $r_n^L = V - V\left(\operatorname{argmax}_{a \in A^L} T_a(M)\right)$, that is we can bound the simple regret of a sequence of actions in A^L rather than only the first action $a(n) \in A$. Thus, using n actions for the exploration of the environment, one can derive a plan of length L (of order $\log n$) with the optimality guarantees of Theorem 5.3.*

4. Discussion

In this section we compare the performance of OLOP with previous algorithms that can be adapted to our framework. This discussion is summarized in Figure 3. We also point out several open questions raised by these comparisons.

Comparison with Zooming Algorithm/HOO: In Kleinberg et al. [2008a] and Bubeck et al. [2009c], the authors consider a very general version of stochastic bandits, where the set of arms $\mathcal{X}$ is a metric space (or even more general spaces in Bubeck et al. [2009c]). When the underlying mean-payoff function is 1-Lipschitz with respect to the metric (again, weaker assumption are derived in Bubeck et al. [2009c]), the authors propose two algorithms, respectively the Zooming Algorithm and HOO, for which they derive performances in terms of either the zooming dimension or the near-optimality dimension. In a metric space, both of these notions coincide, and the corresponding dimension d is defined such that the number of balls of diameter ε required to cover the set of arms that are ε-optimal is of order ε^{-d}. Then, for both algorithms, one obtains a simple regret of order $\widetilde{O}(n^{-1/(d+2)})$ (thanks to Remark 5.3).

Up to minor details, one can see our framework as a A^∞-armed bandit problem, where the mean-payoff function is the sum of discounted rewards. A natural metric ℓ on this space can be defined as follows: For any $a, b \in A^\infty$, $\ell(a, b) = \frac{\gamma^{h(a,b)+1}}{1-\gamma}$, where $h(a, b)$ is the maximum depth $t \geqslant 0$ such that $a_{1:t} = b_{1:t}$. One can very easily check that the sum of discounted reward is 1-Lipschitz with respect to that metric, since $\sum_{t \geqslant 1} \gamma^t |\mu(a_{1:t}) - \mu(b_{1:t})| = \sum_{t \geqslant h(a,b)+1} \gamma^t |\mu(a_{1:t}) - \mu(b_{1:t})| \leqslant \ell(a, b)$. We show now that κ_2, defined by (5.3), is closely related to the near-optimality dimension. Indeed, note that the set aA^∞ can be seen as a ball of diameter $\frac{\gamma^{h(a)+1}}{1-\gamma}$. Thus, from the definition of κ_2, the number of balls of diameter $\frac{\gamma^{h+1}}{1-\gamma}$ required to cover the set of $2\frac{\gamma^{h+1}}{1-\gamma}$-optimal paths is of order of κ^h, which implies that the near-optimality dimension is $d = \frac{\log \kappa}{\log 1/\gamma}$. Thanks to this result, we can see that applying the Zooming Algorithm or HOO in our setting yield a simple

regret bounded as:

$$\mathbb{E}r_n = \widetilde{O}(n^{-1/(d+2)}) = \widetilde{O}(n^{-\frac{\log 1/\gamma}{\log \kappa + 2\log 1/\gamma}}). \tag{5.4}$$

Clearly, this rate is always worse than the ones in Theorem 5.3. In particular, when one takes the supremum over all κ, we find that (5.4) gives the same rate as the one of naive uniform planning in (5.2). This was expected since these algorithms do not use the specific shape of the global reward function (which is the sum of rewards obtained along a sequence) to generalize efficiently across arms. More precisely, they do not consider the fact that a reward sample observed for an arm (or sequence) ab provides strong information about any arm in aA^∞. Actually, the difference between HOO and OLOP is the same as the one between the naive uniform planning and the good one (see Section 2.2).

However, although things are obvious for the case of uniform planning, in the case of OLOP, it is much more subtle to prove that it is indeed possible to collect enough reward samples along sequences $ab, b \in A^*$ to deduce a sharp estimation of $\mu(a)$. Indeed, for uniform planning, if each sequence $ab, b \in A^h$ is chosen once, then one may estimate $\mu(a)$ using K^h reward samples. However in OLOP, since the exploration is expected to focus on promising sequences rather than being uniform, it is much harder to control the number of times a sequence $a \in A^*$ has been played. This difficulty makes the proof of Theorem 5.3 quite intricate compared to the proof of HOO for instance.

Comparison with UCB-AIR: When one knows that there are many near-optimal sequences of actions (i.e. when κ is close to K), then one may be convinced that among a certain number of paths chosen uniformly at random, there exists at least one which is very good with high probability. This idea is exploited by the UCB-AIR algorithm of Wang et al. [2009], designed for infinitely many-armed bandits, where at each round one chooses either to sample a new arm (or sequence in our case) uniformly at random, or to re-sample an arm that has already been explored (using a UCB-like algorithm to choose which one). The regret bound of Wang et al. [2009] is expressed in terms of the probability of selecting an ε-optimal sequence when one chooses the actions uniformly at random. More precisely, the characteristic quantity is β such that this probability is of order of ε^β. Again, one can see that κ_2 is closely related to β. Indeed, our assumption says that the proportion of ε-optimal sequences of actions (with $\varepsilon = 2\frac{\gamma^{h+1}}{1-\gamma}$) is $O(\kappa^h)$, resulting in $\kappa = K\gamma^\beta$. Thanks to this result, we can see that applying UCB-AIR in our setting yield a simple regret bounded as:

$$\mathbb{E}r_n = \begin{cases} \widetilde{O}(n^{-\frac{1}{2}}) & \text{if } \kappa > K\gamma \\ \widetilde{O}(n^{-\frac{1}{1+\beta}}) = \widetilde{O}(n^{-\frac{\log 1/\gamma}{\log K/\kappa + \log 1/\gamma}}) & \text{if } \kappa \leqslant K\gamma \end{cases}$$

As expected, UCB-AIR is very efficient when there is a large proportion of near-optimal paths. Note that UCB-AIR requires the knowledge of β (or equivalently κ).

Figure 3 shows a comparison of the exponents in the simple regret bounds for OLOP, uniform planning, UCB-AIR, and Zooming/HOO (in the case $K\gamma^2 > 1$). We note that the rate for OLOP is better than UCB-AIR when there is a small proportion of near-optimal paths (small κ). Uniform planning is always dominated by OLOP and corresponds to a minimax lower bound for any algorithm. Zooming/HOO are always strictly dominated by OLOP and they do not attain minimax performances.

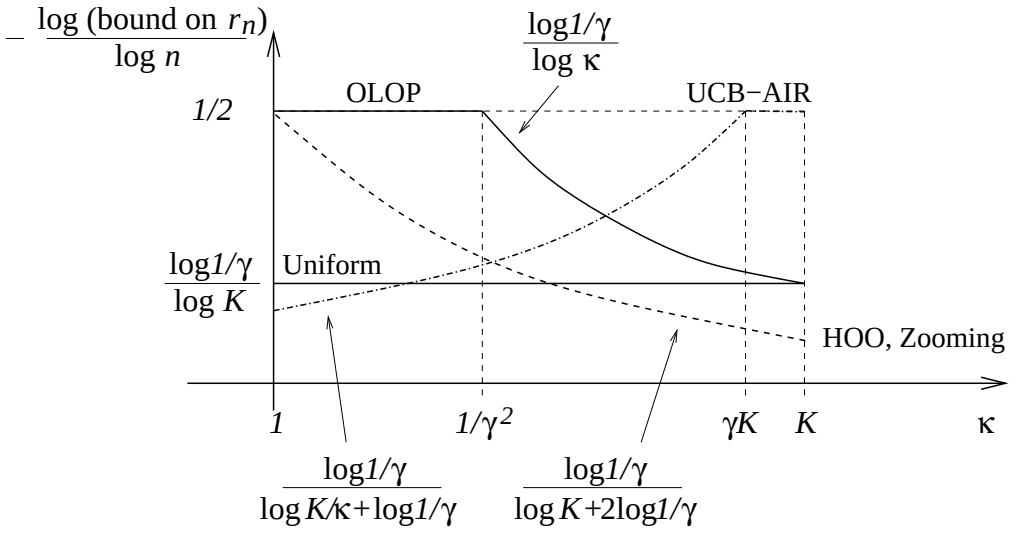

Figure 3: Comparison of the exponent rate of the bounds on the simple regret for OLOP, (good) uniform planning, UCB-AIR, and Zooming/HOO, as a function of $\kappa \in [1, K]$, in the case $K\gamma^2 > 1$.

Comparison with deterministic setting: In Hren and Munos [2008], the authors consider a deterministic version of our framework, precisely they assume that the rewards are a deterministic function of the sequence of actions. Remarkably, in the case $\kappa\gamma^2 > 1$, we obtain the same rate for the simple regret as Hren and Munos [2008]. Thus, in this case, we can say that planning in stochastic environments is not harder than planning in deterministic environments (moreover, note that in deterministic environments there is no distinction between open-loop and closed-loop planning).

Open questions: We identify four important open questions. (i) Is it possible to attain the performances of UCB-AIR when κ is unknown? (ii) Is it possible to improve OLOP if κ is known? (iii) Can we combine the advantages of OLOP and UCB-AIR to derive an exploration strategy with improved rate in intermediate cases (i.e. when $1/\gamma^2 < \kappa < \gamma K$)? (iv) What is a problem-dependent lower bound (in terms of κ or other measures of the environment) in this framework? Obviously these problems are closely related, and the current behavior of the bounds suggests that question (iv) might be tricky. As a side question, note that OLOP requires the knowledge of the time-horizon n, we do not know whether it is possible to obtain the same guarantees with an anytime algorithm.

5. Proofs

5.1. Proof of Theorem 5.1. Let $\varepsilon \in [0, 1/2]$. For $h \geqslant 1$ and $b \in A^h$ we define the environment ν_b as follows. If $a \notin A^h$ then $\nu_b(a) = \delta_0$. If $a \in A^h \setminus \{b\}$ then $\nu_b(a) = \mathrm{Ber}\left(\frac{1-\varepsilon}{2}\right)$. And finally we set $\nu_b(b) = \mathrm{Ber}\left(\frac{1+\varepsilon}{2}\right)$. We also note ν_0 the environment such that if $a \notin A^h$ (respectively $a \in A^h$) then $\nu_0(a) = \delta_0$ (respectively $\nu_0(a) = \mathrm{Ber}\left(\frac{1-\varepsilon}{2}\right)$). Note that, under ν_b, for any $a \in A \setminus \{b_1\}$, we have $V - V(a) = \varepsilon\gamma^h$.

We clearly have

$$\sup_{b\in A^h} \mathbb{E}_{\nu_b} r_n = \sup_{b\in A^h} \varepsilon\gamma^h \mathbb{P}_{\nu_b}(a(n) \neq b_1) \geqslant \gamma^h \varepsilon \left(1 - \frac{1}{K^h} \sum_{b\in A^h} \mathbb{P}_{\nu_b}(a(n) = b_1)\right).$$

Now by Pinsker's inequality we get

$$\mathbb{P}_{\nu_b}(a(n) = b_1) \leqslant \mathbb{P}_{\nu_0}(a(n) = b_1) + \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}_{\nu_0}, \mathbb{P}_{\nu_b}\right)}.$$

Note that

$$\frac{1}{K^h} \sum_{b\in A^h} \mathbb{P}_{\nu_0}(a(n) = b_1) = \frac{1}{K}.$$

Using the chain rule for Kullback Leibler's divergence (see the third step of the proof of Lemma 2.2) we obtain

$$\mathrm{KL}\left(\mathbb{P}_{\nu_0}, \mathbb{P}_{\nu_b}\right) \leqslant \mathrm{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right) \mathbb{E}_{\nu_0} T_b(n).$$

Note also that $\mathrm{KL}\left(\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right) \leqslant \frac{2\varepsilon^2}{1-\varepsilon} \leqslant 4\varepsilon^2$ and thus by the concavity of the square root we obtain:

$$\frac{1}{K^h} \sum_{b\in A^h} \sqrt{\mathbb{E}_{\nu_0} T_b(n)} \leqslant \sqrt{\frac{1}{K^h} \sum_{b\in A^h} \mathbb{E}_{\nu_0} T_b(n)} \leqslant \sqrt{\frac{n}{hK^h}}.$$

So far we proved:

$$\sup_{b\in A^h} \mathbb{E}_{\nu_b} r_n \geqslant \gamma^h \varepsilon \left(1 - \frac{1}{K} - \varepsilon\sqrt{\frac{n}{hK^h}}\right).$$

Taking $\varepsilon = \frac{1}{4}\min(1, \sqrt{hK^h/n})$ yields the lower bound $\frac{1}{16}\gamma^h \min(1, \sqrt{hK^h/n})$. The proof is concluded by taking $h = \log(n\log(1/\gamma^2)/\log n)/\log(1/\gamma^2)$ if $\gamma\sqrt{K} \leqslant 1$ and $h = \log(n\log K/\log n)/\log K$ if $\gamma\sqrt{K} > 1$.

5.2. Proof of Theorem 5.2. First note that, since H is the largest integer such that $HK^H \leqslant n$, it satisfies:

$$\frac{\log n}{\log K} \geqslant H \geqslant \lfloor \frac{\log(n\log K/\log n)}{\log K} \rfloor. \tag{5.5}$$

Let $\widetilde{a} = \arg\max_{a\in A^H} \widehat{V}(a)$ and $a^* \in A^H$ be such that $V(a^*) = V$. Then we have

$$\begin{aligned}
\mathbb{E} r_n &= V - \mathbb{E}\, V(a(n)) \\
&\leqslant V - \mathbb{E}\, V(\widetilde{a}) \\
&\leqslant \frac{\gamma^{H+1}}{1-\gamma} + \sum_{h=1}^{H} \gamma^h (\mu(a^*_{1:h}) - \mathbb{E}\mu(\widetilde{a}_{1:h})) \\
&\leqslant \frac{\gamma^{H+1}}{1-\gamma} + \sum_{h=1}^{H} \gamma^h \mathbb{E}(\widehat{\mu}(a^*_{1:h}) - \widehat{\mu}(\widetilde{a}_{1:h})) + \sum_{h=1}^{H} \gamma^h \mathbb{E}\left(\widehat{\mu}(\widetilde{a}_{1:h}) - \mu(\widetilde{a}_{1:h})\right).
\end{aligned}$$

Now remark that

$$\sum_{h=1}^{H} \gamma^h \mathbb{E}(\widehat{\mu}(a^*_{1:h}) - \widehat{\mu}(\widetilde{a}_{1:h})) = \mathbb{E}\left(\widehat{V}_H(a^*) - \widehat{V}_H(\widetilde{a})\right) \leqslant 0.$$

Moreover by Hoeffding's inequality,

$$\mathbb{E}\max_{a\in A^h} (\widehat{\mu}(a) - \mu(a)) \leqslant \sqrt{\frac{\log K^h}{2K^{H-h}}}.$$

Thus we obtain

$$\mathbb{E}r_n \leqslant \frac{\gamma^{H+1}}{1-\gamma} + \sqrt{\frac{H\log K}{2K^H}}\sum_{h=1}^{H}(\gamma\sqrt{K})^h. \tag{5.6}$$

Now consider the case $\gamma\sqrt{K} > 1$. We have

$$\sqrt{\frac{H\log K}{2K^H}}\sum_{h=1}^{H}(\gamma\sqrt{K})^h = O\bigg(\frac{\sqrt{H}(\gamma\sqrt{K})^H}{\sqrt{K}^H}\bigg) = O(\sqrt{H}\gamma^H).$$

Plugging this into (5.6) and using (5.5), we obtain

$$\mathbb{E}r_n = O(\sqrt{H}\gamma^H) = O\bigg(\sqrt{\log n}\left(\frac{\log n}{n}\right)^{\frac{\log 1/\gamma}{\log K}}\bigg).$$

In the case $\gamma\sqrt{K} = 1$, we have

$$\sqrt{\frac{H\log K}{2K^H}}\sum_{h=1}^{H}(\gamma\sqrt{K})^h = O\bigg(\frac{H^{3/2}}{\sqrt{K}^H}\bigg) = O(H^{3/2}\gamma^H).$$

Plugging this into (5.6) and using (5.5), we obtain:

$$\mathbb{E}r_n = O(H\gamma^H) = O\bigg(\frac{(\log n)^2}{\sqrt{n}}\bigg).$$

Finally, for $\gamma\sqrt{K} < 1$, we have

$$\sqrt{\frac{H\log K}{2K^H}}\sum_{h=1}^{H}(\gamma\sqrt{K})^h = O\bigg(\sqrt{\frac{H}{K^H}}\bigg).$$

Plugging this into (5.6) and using (5.5), we obtain:

$$\mathbb{E}r_n = O\bigg(\sqrt{\frac{H}{K^H}}\bigg) = O\bigg(\frac{\log n}{\sqrt{n}}\bigg).$$

5.3. Proof of Theorem 5.3. The proof of Theorem 5.3 is quite subtle. To present it in a gentle way we adopt a pyramidal proof rather than a pedagogic one. We propose seven lemmas, which we shall not motivate in depth, but prove in details. The precise architecture of the proof is as follows: Lemma 5.1 is a preliminary step, it justifies Remark 5.3. Then Lemma 5.2 underlines the important cases that we have to treat to show that suboptimal arms are not pulled too often. Lemma 5.3 takes care of one of these cases. Then, from Lemma 5.4 to 5.7, each Lemma builds on its predecessor. The main result eventually follows from Lemma 5.1 and 5.7 together with a simple optimization step.

We introduce first a few notations that will be useful. Let $1 \leqslant H \leqslant L$ and $a^* \in A^L$ such that $\Delta(a^*) = 0$. We define now some useful sets for any $1 \leqslant h \leqslant H$ and $0 \leqslant h' < h$;

$$\mathcal{I}_0 = \{\emptyset\},\ \ \mathcal{I}_h = \left\{a \in A^h : \Delta(a) \leqslant \frac{2\gamma^{h+1}}{1-\gamma}\right\},\ \ \mathcal{J}_h = \left\{a \in A^h : a_{1:h-1} \in \mathcal{I}_{h-1} \text{ and } a \notin \mathcal{I}_h\right\}.$$

Note that, from the definition of κ_2, we have that for any $\kappa' > \kappa_2$, there exists a constant C such that for any $h \geqslant 1$,

$$|\mathcal{I}_h| \leqslant C\kappa'. \tag{5.7}$$

Now for $1 \leqslant m \leqslant M$, and $a \in A^t$ with $t \leqslant h$, write

$$\mathcal{P}^a_{h,h'}(m) = \left\{ b \in aA^{h-t} \cap \mathcal{J}_h : T_b(m) \geqslant \frac{8}{\gamma^2}(h+1)^2\gamma^{2(h'-h)} \log M + 1 \right\}.$$

Finally we also introduce the following random variable:

$$\tau^a_{h,h'}(m) = \mathbb{1}\left\{ T_a(m-1) < \frac{8}{\gamma^2}(h+1)^2\gamma^{2(h'-h)} \log M + 1 \leqslant T_a(m) \right\}.$$

LEMMA 5.1. *The following holds true,*

$$r_n \leqslant \frac{2K\gamma^{H+1}}{1-\gamma} + \frac{3K}{M} \sum_{h=1}^{H} \sum_{a \in \mathcal{J}_h} \frac{\gamma^h}{1-\gamma} T_a(M).$$

PROOF. Since $a(n) \in \arg\max_{a \in A} T_a(M)$, we have $T_{a(n)}(M) \geqslant M/K$, and thus:

$$\frac{M}{K}\Big(V - V(a(n))\Big) \leqslant \Big(V - V(a(n))\Big) T_{a(n)}(M) \leqslant \sum_{m=1}^{M} V - V(a^m).$$

Hence, we have, $r_n \leqslant \frac{K}{M}\sum_{m=1}^{M} V - V(a^m)$. Now remark that, for any sequence of actions $a \in A^L$, we have either:

- $a_{1:H} \in \mathcal{I}_H$; which implies $V - V(a) \leqslant \frac{2\gamma^{H+1}}{1-\gamma}$.
- or there exists $1 \leqslant h \leqslant H$ such that $a_{1:h} \in \mathcal{J}_h$; which implies $V - V(a) \leqslant V - V(a_{1:h-1}) + \frac{\gamma^h}{1-\gamma} \leqslant \frac{3\gamma^h}{1-\gamma}$.

Thus we can write:

$$\begin{aligned}
\sum_{m=1}^{M} (V - V(a^m)) &= \sum_{m=1}^{M} (V - V(a^m)) \left(\mathbb{1}\{a^m \in \mathcal{I}_H\} + \mathbb{1}\{\exists 1 \leqslant h \leqslant H : a^m_{1:h} \in \mathcal{J}_h\} \right) \\
&\leqslant \frac{2\gamma^{H+1}}{1-\gamma} M + 3 \sum_{h=1}^{H} \sum_{a \in \mathcal{J}_h} \frac{\gamma^h}{1-\gamma} T_a(M),
\end{aligned}$$

which ends the proof of Lemma 5.1. □

The rest of the proof is devoted to the analysis of the term $\mathbb{E}\sum_{a \in \mathcal{J}_h} T_a(M)$. In the stochastic bandit literature, it is usual to bound the expected number of times a suboptimal action is pulled by the inverse suboptimality (of this action) squared, see for instance Auer et al. [2002] or Bubeck et al. [2009c]. Specialized to our setting, this implies a bound on $\mathbb{E}T_a(M)$, for $a \in \mathcal{J}_h$, of order γ^{-2h}. However, here, we obtain much stronger guarantees, resulting in the faster rates. Namely we show that $\mathbb{E}\sum_{a \in \mathcal{J}_h} T_a(M)$ is of order $(\kappa')^h$ (rather than $(\kappa')^h\gamma^{-2h}$ with previous methods).

The next lemma describes under which circumstances a suboptimal sequence of actions in $\mathcal{J}_h$ can be selected.

LEMMA 5.2. *Let $0 \leqslant m \leqslant M-1$, $1 \leqslant h \leqslant L$ and $a \in \mathcal{J}_h$. If $a^{m+1} \in aA^*$ then it implies that one the three following propositions is true:*

$$\exists 1 \leqslant h' \leqslant L : U_{a^*_{1:h'}}(m) < V, \tag{5.8}$$

or

$$\sum_{t=1}^{h} \gamma^t \widehat{\mu}_{a_{1:t}}(m) \geqslant V(a) + \sum_{t=1}^{h} \gamma^t \sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}}, \tag{5.9}$$

or

$$2\sum_{t=1}^{h}\gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} > \frac{\gamma^{h+1}}{1-\gamma}. \tag{5.10}$$

PROOF. If $a^{m+1} \in aA^*$ then it implies that $U_a(m) \geqslant \inf_{1\leqslant h'\leqslant L} U_{a^*_{1:h'}}(m)$. That is either (5.8) is true or

$$U_a(m) = \sum_{t=1}^{h}\gamma^t\widehat{\mu}_{a_{1:t}}(m) + \gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} + \frac{\gamma^{h+1}}{1-\gamma} \geqslant V.$$

In the latter case, if (5.9) is not satisfied, it implies

$$V(a) + 2\sum_{t=1}^{h}\gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} + \frac{\gamma^{h+1}}{1-\gamma} > V. \tag{5.11}$$

Since $a \in \mathcal{J}_h$ we have $V - V(a) - \frac{\gamma^{h+1}}{1-\gamma} \geqslant \frac{\gamma^{h+1}}{1-\gamma}$ which shows that equation (5.11) implies (5.10) and ends the proof. □

We show now that both equations (5.8) and (5.9) have a vanishing probability of being satisfied.

LEMMA 5.3. *The following holds true, for any* $1 \leqslant h \leqslant L$ *and* $m \leqslant M$*,*

$$\mathbb{P}(\textit{equation (5.8) or (5.9) is true}) \leqslant m(L+h)M^{-4} = \widetilde{O}(M^{-3}).$$

PROOF. Since $V \leqslant \sum_{t=1}^{h}\gamma^t\mu(a^*_{1:t}) + \frac{\gamma^{h+1}}{1-\gamma}$, we have,

$$\begin{aligned}
&\mathbb{P}(\exists\, 1 \leqslant h \leqslant L : U_{a^*_{1:h}}(m) \leqslant V) \\
&\leqslant \mathbb{P}\left(\exists\, 1 \leqslant h \leqslant L : \sum_{t=1}^{h}\gamma^t\left(\widehat{\mu}_{a^*_{1:t}}(m) + \sqrt{\frac{2\log M}{T_{a^*_{1:t}}(m)}}\right) \leqslant \sum_{t=1}^{h}\gamma^t\mu(a^*_{1:t}) \text{ and } T_{a^*_{1:h}}(m) \geqslant 1\right) \\
&\leqslant \mathbb{P}\left(\exists\, 1 \leqslant t \leqslant L : \widehat{\mu}_{a^*_{1:t}}(m) + \sqrt{\frac{2\log M}{T_{a^*_{1:t}}(m)}} \leqslant \mu(a^*_{1:t}) \text{ and } T_{a^*_{1:t}}(m) \geqslant 1\right) \\
&\leqslant \sum_{t=1}^{L}\mathbb{P}\left(\widehat{\mu}_{a^*_{1:t}}(m) + \sqrt{\frac{2\log M}{T_{a^*_{1:t}}(m)}} \leqslant \mu(a^*_{1:t}) \text{ and } T_{a^*_{1:t}}(m) \geqslant 1\right).
\end{aligned}$$

Now we want to apply a concentration inequality to bound this last term. To do it properly we exhibit a martingale and apply the Hoeffding-Azuma inequality for martingale differences (see Theorem 10.1). Let

$$S_j = \min\{s : T_{a^*_{1:t}}(s) = j\},\ j \geqslant 1.$$

If $S_j \leqslant M$, we define $\widetilde{Y}_j = Y_t^{S_j}$, and otherwise $\widetilde{Y}_j$ is an independent random variable with law $\nu(a^*_{1:t})$. We clearly have,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\mu}_{a^*_{1:t}}(m) + \sqrt{\frac{2\log M}{T_{a^*_{1:t}}(m)}} \leqslant \mu(a^*_{1:t}) \text{ and } T_{a^*_{1:t}}(m) \geqslant 1\right) \\
&= \mathbb{P}\left(\frac{1}{T_{a^*_{1:t}}(m)}\sum_{j=1}^{T_{a^*_{1:t}}(m)}\widetilde{Y}_j + \sqrt{\frac{2\log M}{T_{a^*_{1:t}}(m)}} \leqslant \mu(a^*_{1:t}) \text{ and } T_{a^*_{1:t}}(m) \geqslant 1\right) \\
&\leqslant \sum_{u=1}^{m}\mathbb{P}\left(\frac{1}{u}\sum_{j=1}^{u}\widetilde{Y}_j + \sqrt{\frac{2\log M}{u}} \leqslant \mu(a^*_{1:t})\right).
\end{aligned}$$

Now we have to prove that $\widetilde{Y}_j - \mu(a^*_{1:t})$ is martingale differences sequence. This follows via an optional skipping argument, see [Doob, 1953, Chapter VII, Theorem 2.3]. Thus we obtain

$$\mathbb{P}(\text{equation (5.8) is true}) \leqslant \sum_{t=1}^{L}\sum_{u=1}^{m} \exp\left(-2u\frac{2\log M}{u}\right) = LmM^{-4}.$$

The same reasoning gives

$$\mathbb{P}(\text{equation (5.9) is true}) \leqslant mhM^{-4},$$

which concludes the proof. □

The next lemma proves that, if a sequence of actions has already been pulled enough, then equation (5.10) is not satisfied, and thus using lemmas 5.2 and 5.3 we deduce that with high probability this sequence of actions will not be selected anymore. This reasoning is made precise in Lemma 5.5.

LEMMA 5.4. *Let $1 \leqslant h \leqslant L$, $a \in \mathcal{J}_h$ and $0 \leqslant h' < h$. Then equation* (5.10) *is not satisfied if the two following propositions are true:*

$$\forall 0 \leqslant t \leqslant h',\ T_{a_{1:t}}(m) \geqslant \frac{8}{\gamma^2}(h+1)^2\gamma^{2(t-h)}\log M, \tag{5.12}$$

and

$$T_a(m) \geqslant \frac{8}{\gamma^2}(h+1)^2\gamma^{2(h'-h)}\log M. \tag{5.13}$$

PROOF. Assume that (5.12) and (5.13) are true. Then we clearly have:

$$\begin{aligned}
2\sum_{t=1}^{h}\gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} &= 2\mathbb{1}_{h'>0}\sum_{t=1}^{h'}\gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} + 2\sum_{t=h'+1}^{h}\gamma^t\sqrt{\frac{2\log M}{T_{a_{1:t}}(m)}} \\
&\leqslant \frac{\gamma^{h+1}}{h+1}h' + \frac{\gamma^{h+1}}{h+1}\sum_{t=h'+1}^{h}\gamma^{t-h'} \\
&\leqslant \frac{\gamma^{h+1}}{h+1}\left(h' + \frac{\gamma}{1-\gamma}\right) \\
&\leqslant \frac{\gamma^{h+1}}{1-\gamma},
\end{aligned}$$

which proves the result. □

LEMMA 5.5. *Let $1 \leqslant h \leqslant L$, $a \in \mathcal{J}_h$ and $0 \leqslant h' < h$. Then $\tau^a_{h,h'}(m+1) = 1$ implies that either equation* (5.8) *or* (5.9) *is satisfied or the following proposition is true:*

$$\exists 0 \leqslant t \leqslant h' : |\mathcal{P}^{a_{1:t}}_{h,h'}(m)| < \gamma^{2(t-h')}. \tag{5.14}$$

PROOF. If $\tau^a_{h,h'}(m+1) = 1$ then it means that $a^{m+1} \in aA^*$ and (5.13) is satisfied. By Lemma 5.2 this implies that either (5.8), (5.9) or (5.10) is true and (5.13) is satisfied. Now by Lemma 5.4 this implies that (5.8) is true or (5.9) is true or (5.12) is false. We now prove that if (5.14) is not satisfied then (5.12) is true, which clearly ends the proof. This follows from: For any $0 \leqslant t \leqslant h'$,

$$T_{a_{1:t}}(m) = \sum_{b \in a_{1:t}A^{h-t}} T_b(m) \geqslant \gamma^{2(t-h')}\frac{8}{\gamma^2}(h+1)^2\gamma^{2(h'-h)}\log M = \frac{8}{\gamma^2}(h+1)^2\gamma^{2(t-h)}\log M.$$

□

The next lemma is the key step of our proof. Intuitively, using lemmas 5.2 and 5.5, we have a good control on sequences for which equation (5.14) is satisfied. Note that (5.14) is a property which depends on sub-sequences of a from length 1 to h'. In the following proof we will iteratively "drop" all sequences which do not satisfy (5.14) from length t onwards, starting from $t = 1$. Then, on the remaining sequences, we can apply Lemma 5.5.

LEMMA 5.6. *Let $1 \leqslant h \leqslant L$ and $0 \leqslant h' < h$. Then the following holds true,*

$$\mathbb{E}|\mathcal{P}^{\emptyset}_{h,h'}(M)| = \widetilde{O}\left(\gamma^{-2h'}\mathbb{1}_{h'>0}\sum_{t=0}^{h'}(\gamma^2\kappa')^t + (\kappa')^h M^{-2}\right).$$

PROOF. Let $h' \geqslant 1$ and $0 \leqslant s \leqslant h'$. We introduce the following random variables:

$$m^a_s = \min\left(M, \min\left\{m \geqslant 0 : |\mathcal{P}^a_{h,h'}(m)| \geqslant \gamma^{2(s-h')}\right\}\right).$$

We will prove recursively that,

$$|\mathcal{P}^{\emptyset}_{h,h'}(m)| \leqslant \sum_{t=0}^{s}\gamma^{2(t-h')}|\mathcal{I}_t| + \sum_{a\in\mathcal{I}_s}\left|\mathcal{P}^a_{h,h'} \setminus \cup_{t=0}^{s}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right|. \tag{5.15}$$

The result is true for $s = 0$ since $\mathcal{I}_0 = \{\emptyset\}$ and by definition of $m^{\emptyset}_0$,

$$|\mathcal{P}^{\emptyset}_{h,h'}(m)| \leqslant \gamma^{-2h'} + |\mathcal{P}^{\emptyset}_{h,h'}(m) \setminus \mathcal{P}^{\emptyset}_{h,h'}(m^{\emptyset}_0)|.$$

Now let us assume that the result is true for $s < h'$. We have:

$$\begin{aligned}
\sum_{a\in\mathcal{I}_s}\left|\mathcal{P}^a_{h,h'}(m) \setminus \cup_{t=0}^{s}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right| &= \sum_{a\in\mathcal{I}_{s+1}}\left|\mathcal{P}^a_{h,h'}(m) \setminus \cup_{t=0}^{s}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right| \\
&\leqslant \sum_{a\in\mathcal{I}_{s+1}}\gamma^{2(s+1-h')} + \left|\mathcal{P}^a_{h,h'}(m) \setminus \cup_{t=0}^{s+1}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right| \\
&= \gamma^{2(s+1-h')}|\mathcal{I}_{s+1}| + \sum_{a\in\mathcal{I}_{s+1}}\left|\mathcal{P}^a_{h,h'}(m) \setminus \cup_{t=0}^{s+1}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right|,
\end{aligned}$$

which ends the proof of (5.15). Thus we proved (by taking $s = h'$ and $m = M$):

$$\begin{aligned}
|\mathcal{P}^{\emptyset}_{h,h'}(M)| &\leqslant \sum_{t=0}^{h'}\gamma^{2(t-h')}|\mathcal{I}_t| + \sum_{a\in\mathcal{I}_{h'}}\left|\mathcal{P}^a_{h,h'}(M) \setminus \cup_{t=0}^{s+1}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right| \\
&= \sum_{t=0}^{h'}\gamma^{2(t-h')}|\mathcal{I}_t| + \sum_{a\in\mathcal{J}_h}\left|\mathcal{P}^a_{h,h'}(M) \setminus \cup_{t=0}^{s+1}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right|
\end{aligned}$$

Now, for any $a \in \mathcal{J}_h$, let $\widetilde{m} = \max_{0\leqslant t\leqslant h'} m^{a_{1:t}}_t$. Note that for $m \geqslant \widetilde{m}$, equation (5.14) is not satisfied. Thus we have

$$\begin{aligned}
\left|\mathcal{P}^a_{h,h'} \setminus \cup_{t=0}^{s+1}\mathcal{P}^{a_{1:t}}_{h,h'}(m^{a_{1:t}}_t)\right| = \sum_{m=\widetilde{m}}^{M-1}\tau^a_{h,h'}(m+1) &= \sum_{m=0}^{M-1}\tau^a_{h,h'}(m+1)\ \mathbb{1}\{\text{(5.14) is not satisfied}\} \\
&\leqslant \sum_{m=0}^{M-1}\tau^a_{h,h'}(m+1)\ \mathbb{1}\{\text{(5.8) or (5.9) is satisfied}\}.
\end{aligned}$$

where the last inequality results from Lemma 5.5. Hence, we proved:

$$|\mathcal{P}^{\emptyset}_{h,h'}| \leqslant \sum_{t=0}^{h'} \gamma^{2(t-h')}|\mathcal{I}_t| + \sum_{m=0}^{M-1}\sum_{a\in\mathcal{J}_h} \mathbb{1}\{\text{(5.8) or (5.9) is satisfied}\}.$$

Taking the expectation, using (5.7) and applying Lemma 5.3 yield the claimed bound for $h' \geqslant 1$.

Now for $h' = 0$ we need a modified version of Lemma 5.5. Indeed in this case one can directly prove that $\tau^a_{h,0}(m+1) = 1$ implies that either equation (5.8) or (5.9) is satisfied (this follows from the fact that $\tau^a_{h,0}(m+1) = 1$ always imply that (5.12) is true for $h' = 0$). Thus we obtain:

$$|\mathcal{P}^{\emptyset}_{h,h'}| = \sum_{m=0}^{M-1}\sum_{a\in\mathcal{J}_h} \tau^a_{h,0}(m+1) \leqslant \sum_{m=0}^{M-1}\sum_{a\in\mathcal{J}_h} \mathbb{1}\{\text{(5.8) or (5.9) is satisfied}\}.$$

Taking the expectation and applying Lemma 5.3 yield the claimed bound for $h' = 0$ and ends the proof. □

LEMMA 5.7. *Let $1 \leqslant h \leqslant L$. The following holds true,*

$$\mathbb{E}\sum_{a\in\mathcal{J}_h} T_a(M) = \widetilde{O}\left(\gamma^{-2h}\sum_{h'=1}^{h}(\gamma^2\kappa')^{h'} + (\kappa')^h(1+\gamma^{-2h}M^{-2})\right).$$

PROOF. We have the following computations:

$$\begin{aligned}
\sum_{a\in\mathcal{J}_h} T_a(M) &= \sum_{a\in\mathcal{J}_h\setminus\mathcal{P}^{\emptyset}_{h,h-1}} T_a(M) + \sum_{h'=1}^{h-1}\sum_{a\in\mathcal{P}^{\emptyset}_{h,h'}\setminus\mathcal{P}^{\emptyset}_{h,h'-1}} T_a(M) + \sum_{a\in\mathcal{P}^{\emptyset}_{h,0}} T_a(M) \\
&\leqslant \frac{8}{\gamma^2}(h+1)^2\gamma^{2(h-1-h)}|\mathcal{J}_h| + \sum_{h'=1}^{h-1}\frac{8}{\gamma^2}(h+1)^2\gamma^{2(h'-1-h)}\log M|\mathcal{P}^{\emptyset}_{h,h'}| + M|\mathcal{P}^{\emptyset}_{h,0}| \\
&= \widetilde{O}\left((\kappa')^h + \gamma^{-2h}\sum_{h'=1}^{h-1}\gamma^{2h'}|\mathcal{P}^{\emptyset}_{h,h'}| + M|\mathcal{P}^{\emptyset}_{h,0}|\right).
\end{aligned}$$

Taking the expectation and applying the bound of Lemma 5.6 gives the claimed bound. □

Thus by combining Lemma 5.1 and 5.7 we obtain for $\kappa'\gamma^2 \leqslant 1$:

$$\mathbb{E}r_n = \widetilde{O}\left(\gamma^H + \gamma^{-H}M^{-1} + (\kappa')^H\gamma^{-H}M^{-3}\right),$$

and for $\kappa'\gamma^2 > 1$:

$$\mathbb{E}r_n = \widetilde{O}\left(\gamma^H + (\kappa'\gamma)^H M^{-1} + (\kappa')^H\gamma^{-H}M^{-3}\right).$$

Thus in the case $\kappa'\gamma^2 \leqslant 1$, taking $H = \lfloor \log M/(2\log 1/\gamma)\rfloor$ yields the claimed bound; while for $\kappa'\gamma^2 > 1$ we take $H = \lfloor \log M/\log\kappa'\rfloor$. Note that in both cases we have $H \leqslant L$ (as it was required at the beginning of the analysis).

CHAPTER 6

Pure Exploration in Multi-Armed Bandits

We consider the framework of stochastic multi-armed bandit problems and study the possibilities and limitations of forecasters that perform an on-line exploration of the arms. A forecaster is assessed in terms of its simple regret, a regret notion that captures the fact that exploration is only constrained by the number of available rounds (not necessarily known in advance), in contrast to the case when the cumulative regret is considered and when exploitation needs to be performed at the same time. We believe that this performance criterion is suited to situations when the cost of pulling an arm is expressed in terms of resources rather than rewards. We discuss the links between the simple and the cumulative regret. The main result is that the required exploration–exploitation trade-offs are qualitatively different, in view of a general lower bound on the simple regret in terms of the cumulative regret.

Contents

This chapter is a joint work with Rémi Munos and Gilles Stoltz. It is based on the extended version Bubeck et al. [2009b] (currently under submission) of Bubeck et al. [2009a] which appeared in the Proceedings of the 20th International Conference on Algorithmic Learning Theory.

1. Introduction

Learning processes usually face an exploration versus exploitation dilemma, since they have to get information on the environment (exploration) to be able to take good actions (exploitation). A key example is the multi-armed bandit problem described in Chapter 2. The usual assessment criterion of a forecaster is given by its cumulative regret and typical good forecasters, like UCB of Auer et al. [2002], trade off between exploration and exploitation.

Our setting is as follows. The forecaster may sample the arms a given number of times n (not necessarily known in advance) and is then asked to output a recommended arm. He is evaluated by his simple regret, that is, the difference between the average payoff of the best arm and the average payoff obtained by his recommendation. The distinguishing feature from the classical multi-armed bandit problem is that the exploration phase and the evaluation phase are separated. We now illustrate why this is a natural framework for numerous applications.

Historically, the first occurrence of multi-armed bandit problems was given by medical trials. In the case of a severe disease, ill patients only are included in the trial and the cost of picking the wrong treatment is high. It is important to minimize the cumulative regret, since the test and cure phases coincide. However, for cosmetic products, there exists a test phase separated from the commercialization phase, and one aims at minimizing the regret of the commercialized product rather than the cumulative regret in the test phase, which is irrelevant. (Here, several formulæ for a cream are considered and some quantitative measurement, like skin moisturization, is performed.)

The pure exploration problem addresses the design of strategies making the best possible use of available numerical resources (e.g., as CPU time) in order to optimize the performance of some decision-making task. That is, it occurs in situations with a preliminary exploration phase in which costs are not measured in terms of rewards but rather in terms of resources, that come in limited budget.

A motivating example concerns recent works on computer-go (e.g., the MoGo program of Gelly et al. [2006]). A given time, i.e., a given amount of CPU times is given to the player to explore the possible outcome of sequences of plays and output a final decision. An efficient exploration of the search space is obtained by considering a hierarchy of forecasters minimizing some cumulative regret – see, for instance, the UCT strategy of Kocsis and Szepesvari [2006] and the BAST strategy of Coquelin and Munos [2007]. However, the cumulative regret does not seem to be the right way to base the strategies on, since the simulation costs are the same for exploring all options, bad and good ones. This observation was actually the starting point of the notion of simple regret and of this work.

A final related example is the maximization of some function f, observed with noise. Whenever evaluating f at a point is costly (e.g., in terms of numerical or financial costs), the issue is to choose as adequately as possible where to query the value of this function in order to have a good approximation to the maximum. The pure exploration problem considered here addresses exactly the design of adaptive exploration strategies making the best use of available resources in order to make the most precise prediction once all resources are consumed.

As a remark, it also turns out that in all examples considered above, we may impose the further restriction that the forecaster ignores ahead of time the amount of available resources (time, budget, or the number of patients to be included) – that is, we seek for anytime performance. We refer the reader to Chapter 7 for an in-depth study of strategies making use of the time horizon.

We end this introduction with an overview of the literature. The problem of pure exploration presented above was referred to as "budgeted multi-armed bandit problem" in the open problem Madani et al. [2004] (where, however, another notion of regret than simple regret is considered). Schlag [2006] solves the pure exploration problem in a minimax sense for the case of two arms only and rewards given by probability distributions over $[0, 1]$. Even-Dar et al. [2002] and Mannor and Tsitsiklis [2004] consider a related setting where forecasters perform exploration during a random number of rounds T and aim at identifying an ε–best arm. They study the possibilities and limitations of policies achieving this goal with overwhelming $1 - \delta$ probability and indicate in particular upper and lower bounds on (the expectation of) T.

Parameters: K unknown probability distributions for the rewards of the arms, $\nu_1, \ldots, \nu_K$.

For each round $t = 1, 2, \ldots,$

(1) the forecaster chooses $I_t \in \{1, \ldots, K\}$;
(2) the environment draws the reward Y_t for that action (also denoted by $X_{I_t, T_{I_t}(t)}$ with the notation introduced in the text);
(3) the forecaster outputs a recommendation $J_t \in \{1, \ldots, K\}$;
(4) if the environment sends a stopping signal, then the game takes an end; otherwise, the next round starts.

Figure 1: The anytime pure exploration problem for multi-armed bandits.

2. Problem setup, notation

We consider a sequential decision problem given by stochastic multi-armed bandits. $K \geqslant 2$ arms, denoted by $i = 1, \ldots, K$, are available and the i–th of them is parameterized by a fixed (unknown) probability distribution ν_i over $[0, 1]$, with expectation denoted by μ_i. At those rounds when it is pulled, its associated reward is drawn at random according to ν_i, independently of all previous rewards. For each arm i and all time rounds $n \geqslant 1$, we denote by $T_i(n)$ the number of times arm i was pulled from rounds 1 to n, and by $X_{i,1}, X_{i,2}, \ldots, X_{i,T_i(n)}$ the sequence of associated rewards.

The forecaster has to deal simultaneously with two tasks, a primary one and an associated one.

The associated task consists in exploration, i.e., the forecaster should indicate at each round t the arm I_t to be pulled, based on past rewards (so that I_t is a random variable). Then the forecaster gets to see the associated reward Y_t, also denoted by $X_{I_t, T_{I_t}(t)}$ with the notation above. The sequence of random variables (I_t) is referred to as an allocation strategy.

The primary task is to output at the end of each round t a recommendation J_t to be used in a one-shot instance if/when the environment sends some stopping signal meaning that the exploration phase is over. The sequence of random variables (J_t) is referred to as a recommendation strategy. In total, a forecaster is given by an allocation and a recommendation strategy.

Figure 1 summarizes the description of the sequential game and points out that the information available to the forecaster for choosing I_t, respectively J_t, is formed by the $X_{i,s}$ for $i = 1, \ldots, K$ and $s = 1, \ldots, T_i(t-1)$, respectively, $s = 1, \ldots, T_i(t)$. Note that we also allow the forecaster to use an external randomization in the definition of I_t and J_t.

As we are only interested in the performances of the recommendation strategy (J_t), we call this problem the pure exploration problem for multi-armed bandits and evaluate the forecaster through its simple regret, defined as follows. First, we denote by

$$\mu^* = \mu_{i^*} = \max_{i=1,\ldots,K} \mu_i$$

the expectation of the rewards of the best arm i^* (a best arm, if there are several of them with same maximal expectation). A useful notation in the sequel is the gap $\Delta_i = \mu^* - \mu_i$ between the maximal expected reward and the one of the i–th arm; as well as the minimal gap

$$\Delta = \min_{i:\Delta_i > 0} \Delta_i \, .$$

Now, the simple regret at round n equals the regret on a one-shot instance of the game for the recommended arm J_n, that is, put more formally,

$$r_n = \mu^* - \mu_{J_n} = \Delta_{J_n} \, .$$

A quantity of related interest is the cumulative regret at round n, which is defined as

$$R_n = \sum_{t=1}^{n} \mu^* - \mu_{I_t} \, .$$

A popular treatment of the multi-armed bandit problems is to construct forecasters ensuring that $\mathbb{E}R_n = o(n)$, see Chapter 2. The quantity $r'_t = \mu^* - \mu_{I_t}$ is sometimes called instantaneous regret. It differs from the simple regret r_t and in particular, $R_n = r'_1 + \ldots + r'_n$ is in general not equal to $r_1 + \ldots + r_n$. Theorem 6.1, among others, will however indicate some connections between r_n and R_n.

Goal and structure of the chapter. We study the links between the simple and the cumulative regret. Intuitively, an efficient allocation strategy for the simple regret should rely on some exploration–exploitation trade-off. Our main contribution (Theorem 6.1, Section 3) is a lower bound on the simple regret in terms of the cumulative regret suffered in the exploration phase, showing that the trade-off involved in the minimization of the simple regret is somewhat different from the one for the cumulative regret. The full consequences of this result are derived in Chapter 7 where we propose new strategies specifically designed for the simple regret. In this chapter, and precisely in Sections 4 and 5, we show how, despite all, one can fight against this negative result. For instance, some strategies designed for the cumulative regret can outperform (for moderate values of n) strategies with exponential rates of convergence for their simple regret. Finally in Section 6 we briefly investigate the $\mathcal{X}$-armed bandit presented in Chapter 4. In this setting we use the simple regret as a tool to characterize the topological spaces $\mathcal{X}$ for which it is possible to have a sublinear cumulative regret.

3. The smaller the cumulative regret, the larger the simple regret

It is immediate that for well-chosen recommendation strategies, the simple regret can be upper bounded in terms of the cumulative regret. For instance, the strategy that at time n recommends arm i with probability $T_i(n)/n$ (recall that we allow the forecaster to use an external randomization) ensures that the simple regret satisfies $\mathbb{E}r_n = \mathbb{E}R_n/n$. Therefore, upper bounds on $\mathbb{E}R_n$ lead to upper bounds on $\mathbb{E}r_n$.

We show here that, conversely, upper bounds on $\mathbb{E}R_n$ also lead to lower bounds on $\mathbb{E}r_n$: the smaller the guaranteed upper bound on $\mathbb{E}R_n$, the larger the lower bound on $\mathbb{E}r_n$, no matter what the recommendation strategy is.

This is interpreted as a variation of the "classical" trade-off between exploration and exploitation. Here, while the recommendation strategy (J_n) relies only on the exploitation of the results of the preliminary exploration phase, the design of the allocation strategy (I_t) consists in an efficient exploration of the arms. To guarantee this efficient exploration, past payoffs of the arms have to be considered and thus, even in the exploration phase, some exploitation is needed. Theorem 6.1 and its corollaries aim at quantifying the needed respective amount of exploration and exploitation. In particular, to have an asymptotic optimal rate of decrease for the simple regret, each arm should be sampled a linear number of times, while for the cumulative regret, it is known that the forecaster should not do so more than a logarithmic number of times on the suboptimal arms.

Formally, our main result is as follows. It is strong in the sense that we get lower bounds for *all* possible *sets* of Bernoulli distributions $\{\nu_1, \ldots, \nu_K\}$ over the rewards. Note that the stated result requires in particular that there is a unique best arm.

THEOREM 6.1 (Main result). *For any forecaster (i.e., for any pair of allocation and recommendation strategies) and any function $\varepsilon : \{1, 2, \ldots\} \to \mathbb{R}$ such that*

> *for all (Bernoulli) distributions $\nu_1, \ldots, \nu_K$ on the rewards, there exists a constant $C \geqslant 0$ with $\mathbb{E}R_n \leqslant C\,\varepsilon(n)$,*

the following holds true:

> *for all sets of $K \geqslant 3$ distinct Bernoulli distributions on the rewards, with parameters different from 1, there exists a constant $D \geqslant 0$ and an ordering $\nu_1, \ldots, \nu_K$ of the considered distributions such that*
>
> $$\mathbb{E}r_n \geqslant e^{-D\varepsilon(n)} .$$

COROLLARY 6.1 (General distribution-dependent lower bound). *For any forecaster, and any set of $K \geqslant 3$ distinct, Bernoulli distributions on the rewards, with parameters different from 1, there exists $\gamma \geqslant 0$ such that, up to the choice of a good ordering of the considered distributions,*

$$\mathbb{E}r_n \geqslant e^{-\gamma n} .$$

Theorem 6.1 is proved below and Corollary 6.1 follows from the fact that the cumulative regret is always bounded by n. To get further the point of the theorem, one should keep in mind that the typical (distribution-dependent) rate of growth of the cumulative regret of good algorithms, e.g., UCB described in Section 2.2 (and which we recall in Figure 2), is $\varepsilon(n) = \ln n$, see Theorem 2.2. This, as asserted in Theorem 2.7, is the optimal rate. But a recommendation strategy based on such allocation strategy is bound to suffer a simple regret that decreases at best polynomially fast. The next result follows from noting that UCB (with exploration parameter α) actually achieves a cumulative regret bounded by a large enough distribution-dependent constant times $\varepsilon(n) = \alpha \ln n$.

COROLLARY 6.2 (Distribution-dependent lower bound for UCB). *The allocation strategy (I_t) given by UCB ensures that for any recommendation strategy (J_t) and all sets of $K \geqslant 3$ distinct, Bernoulli distributions on the rewards, with parameters different from 1, there exists $\gamma \geqslant 0$ (independent of α) such that, up to the choice of a good ordering of the considered distributions,*

$$\mathbb{E}r_n \geqslant n^{-\gamma\alpha} .$$

PROOF. The intuitive version of the proof of Theorem 6.1 is as follows. The basic idea is to consider a tie case when the best and worst arms have zero empirical means; it happens often enough (with a probability at least exponential in the number of times we pulled these arms) and results in the forecaster basically having to pick another arm and suffering some regret. Permutations are used to control the case of untypical or naive forecasters that would despite all pull an arm with zero empirical mean, since they force a situation when those forecasters choose the worst arm instead of the best one.

Formally, we fix the forecaster (a pair of allocation and recommendation strategies) and a corresponding function ε such that the assumption of the theorem is satisfied. We denote by $\boldsymbol{p}_n = (p_{1,n}, \ldots, p_{K,n})$ the probability distribution from which J_n is drawn at random thanks to an auxiliary distribution. Note that $\boldsymbol{p}_n$ is a random vector which depends on $I_1, \ldots, I_n$ as well as on the obtained rewards $Y_1, \ldots, Y_n$. We consider below a set of $K \geqslant 3$ distinct Bernoulli distributions, satisfying the conditions of the theorem; actually, we only use below that their parameters are (up to a first ordering) such that $1 > \mu_1 > \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_K \geqslant 0$ and $\mu_2 > \mu_K$ (thus, $\mu_2 > 0$).

Step 0 introduces another layer of notation. The latter depends on permutations σ of $\{1, \ldots, K\}$. To have a gentle start, we first describe the notation when the permutation is the identity, $\sigma = \mathrm{id}$. We denote by $\mathbb{P}$ and $\mathbb{E}$ the probability and expectation with respect to the original K-tuple $\nu_1, \ldots, \nu_K$ of distributions over the arms. For $i = 1$ (respectively, $i = K$), we denote by $\mathbb{P}_{i,\mathrm{id}}$ and $\mathbb{E}_{i,\mathrm{id}}$ the probability and expectation with respect to the K-tuples formed by $\delta_0, \nu_2, \ldots, \nu_K$ (respectively, $\delta_0, \nu_2, \ldots, \nu_{K-1}, \delta_0$), where δ_0 denotes the Dirac measure on 0.

For a given permutation σ, we consider a similar notation up to a reordering, as follows. $\mathbb{P}_\sigma$ and $\mathbb{E}_\sigma$ refer to the probability and expectation with respect to the K-tuple of distributions over the arms formed by the $\nu_{\sigma^{-1}(1)}, \ldots, \nu_{\sigma^{-1}(K)}$. Note in particular that the i–th best arm is located in the $\sigma(i)$–th position. Now, we denote for $i = 1$ (respectively, $i = K$) by $\mathbb{P}_{i,\sigma}$ and $\mathbb{E}_{i,\sigma}$ the probability and expectation with respect to the K-tuple formed by the $\nu_{\sigma^{-1}(i)}$, except that we replaced the best of them, located in the $\sigma(1)$–th position, by a Dirac measure on 0 (respectively, the best and worst of them, located in the $\sigma(1)$–th and $\sigma(K)$–th positions, by Dirac measures on 0). We provide now a proof in six steps.

Step 1 lower bounds the quantity of interest by an average the maximum of the simple regrets obtained by reordering,

$$\max_\sigma \ \mathbb{E}_\sigma r_n \geqslant \frac{1}{K!} \sum_\sigma \ \mathbb{E}_\sigma r_n \geqslant \frac{\mu_1 - \mu_2}{K!} \sum_\sigma \mathbb{E}_\sigma \big[1 - p_{\sigma(1),n} \big] \ ,$$

where we used that under $\mathbb{P}_\sigma$, the index of the best arm is $\sigma(1)$ and the minimal regret for playing any other arm is at least $\mu_1 - \mu_2$.

Step 2 rewrites each term of the sum over σ as the product of three simple terms. We use first that $\mathbb{P}_{1,\sigma}$ is the same as $\mathbb{P}_\sigma$, except that it ensures that arm $\sigma(1)$ has zero reward throughout. Denoting by

$$C_{i,n} = \sum_{t=1}^{T_i(n)} X_{i,t}$$

the cumulative reward of the i–th arm till round n, one then gets

$$\begin{aligned} \mathbb{E}_\sigma \big[1 - p_{\sigma(1),n} \big] &\geqslant \mathbb{E}_\sigma \Big[\big(1 - p_{\sigma(1),n} \big) \mathbb{1}_{\{C_{\sigma(1),n} = 0\}} \Big] \\ &= \mathbb{E}_\sigma \Big[1 - p_{\sigma(1),n} \ \big| \ C_{\sigma(1),n} = 0 \Big] \times \mathbb{P}_\sigma \big\{ C_{\sigma(1),n} = 0 \big\} \\ &= \mathbb{E}_{1,\sigma} \big[1 - p_{\sigma(1),n} \big] \ \mathbb{P}_\sigma \big\{ C_{\sigma(1),n} = 0 \big\} \ . \end{aligned}$$

Second, iterating the argument from $\mathbb{P}_{1,\sigma}$ to $\mathbb{P}_{K,\sigma}$,

$$\begin{aligned} \mathbb{E}_{1,\sigma} \big[1 - p_{\sigma(1),n} \big] &\geqslant \mathbb{E}_{1,\sigma} \Big[1 - p_{\sigma(1),n} \ \big| \ C_{\sigma(K),n} = 0 \Big] \ \mathbb{P}_{1,\sigma} \big\{ C_{\sigma(K),n} = 0 \big\} \\ &= \mathbb{E}_{K,\sigma} \big[1 - p_{\sigma(1),n} \big] \ \mathbb{P}_{1,\sigma} \big\{ C_{\sigma(K),n} = 0 \big\} \end{aligned}$$

and therefore,

$$\mathbb{E}_\sigma \big[1 - p_{\sigma(1),n} \big] \geqslant \mathbb{E}_{K,\sigma} \big[1 - p_{\sigma(1),n} \big] \ \mathbb{P}_{1,\sigma} \big\{ C_{\sigma(K),n} = 0 \big\} \ \mathbb{P}_\sigma \big\{ C_{\sigma(1),n} = 0 \big\} \ . \tag{6.1}$$

Step 3 deals with the second term in the right-hand side of (6.1),

$$\mathbb{P}_{1,\sigma} \big\{ C_{\sigma(K),n} = 0 \big\} = \mathbb{E}_{1,\sigma} \Big[\big(1 - \mu_K \big)^{T_{\sigma(K)}(n)} \Big] \geqslant \big(1 - \mu_K \big)^{\mathbb{E}_{1,\sigma} T_{\sigma(K)}(n)} \ ,$$

where the equality can be seen by conditioning on $I_1, \ldots, I_n$ and then taking the expectation, whereas the inequality is a consequence of Jensen's inequality. Now, the expected number of times the suboptimal arm $\sigma(K)$ is pulled under $\mathbb{P}_{1,\sigma}$ (for which $\sigma(2)$ is the optimal arm) is bounded by

the regret, by the very definition of the latter: $(\mu_2 - \mu_K)\,\mathbb{E}_{1,\sigma} T_{\sigma(K)}(n) \leqslant \mathbb{E}_{1,\sigma} R_n$. Since by hypothesis (and by taking the maximum of $K!$ values), there exists a constant C such that for all σ, $\mathbb{E}_{1,\sigma} R_n \leqslant C\,\varepsilon(n)$, we finally get

$$\mathbb{P}_{1,\sigma}\big\{C_{\sigma(K),n} = 0\big\} \geqslant (1-\mu_K)^{C\varepsilon(n)/(\mu_2-\mu_K)} \ .$$

Step 4 lower bounds the third term in the right-hand side of (6.1) as

$$\mathbb{P}_{\sigma}\big\{C_{\sigma(1),n} = 0\big\} \geqslant (1-\mu_1)^{C\varepsilon(n)/\mu_2} \ .$$

We denote by $W_n = (I_1, Y_1, \ldots, I_n, Y_n)$ the history of pulled arms and obtained payoffs up to time n. What follows is reminiscent of the techniques used in Mannor and Tsitsiklis [2004]. We are interested in realizations $w_n = (i_1, y_1, \ldots, i_n, y_n)$ of the history such that whenever $\sigma(1)$ was played, it got a null reward. (We denote above by $t_j(t)$ is the realization of $T_j(t)$ corresponding to w_n, for all j and t.) The likelihood of such a w_n under $\mathbb{P}_\sigma$ is $(1-\mu_1)^{t_{\sigma(1)}(n)}$ times the one under $\mathbb{P}_{1,\sigma}$. Thus,

$$\begin{aligned}\mathbb{P}_{\sigma}\big\{C_{\sigma(1),n} = 0\big\} &= \sum \mathbb{P}_\sigma\left\{W_n = w_n\right\} \\ &= \sum (1-\mu_1)^{t_{\sigma(1)}(n)}\ \mathbb{P}_{1,\sigma}\left\{W_n = w_n\right\} = \mathbb{E}_{1,\sigma}\left[(1-\mu_1)^{T_{\sigma(1)}(n)}\right] \ ,\end{aligned}$$

where the sums are over those histories w_n such that the realizations of the payoffs obtained by the arm $\sigma(1)$ equal $x_{\sigma(1),s} = 0$ for all $s = 1, \ldots, t_{\sigma(1)}(n)$. The argument is concluded as before, first by Jensen's inequality and then, by using that $\mu_2\,\mathbb{E}_{1,\sigma} T_{\sigma(1)}(n) \leqslant \mathbb{E}_{1,\sigma} R_n \leqslant C\,\varepsilon(n)$ by definition of the regret and the hypothesis put on its control.

Step 5 resorts to a symmetry argument to show that as far as the first term of the right-hand side of (6.1) is concerned,

$$\sum_\sigma \mathbb{E}_{K,\sigma}\Big[1 - p_{\sigma(1),n}\Big] \geqslant \frac{K!}{2}.$$

Since $\mathbb{P}_{K,\sigma}$ only depends on $\sigma(2), \ldots, \sigma(K-1)$, we denote by $\mathbb{P}^{\sigma(2),\ldots,\sigma(K-1)}$ the common value of these probability distributions when $\sigma(1)$ and $\sigma(K)$ vary (and a similar notation for the associated expectation). We can thus group the permutations σ two by two according to these $(K-2)$–tuples, one of the two permutations being defined by $\sigma(1)$ equal to one of the two elements of $\{1, \ldots, K\}$ not present in the $(K-2)$–tuple, and the other one being such that $\sigma(1)$ equals the other such element. Formally,

$$\begin{aligned}\sum_\sigma \mathbb{E}_{K,\sigma} p_{\sigma(1),n} &= \sum_{j_2,\ldots,j_{K-1}} \mathbb{E}^{j_2,\ldots,j_{K-1}}\left[\sum_{j\in\{1,\ldots,K\}\setminus\{j_2,\ldots,j_{K-1}\}} p_{j,n}\right] \\ &\leqslant \sum_{j_2,\ldots,j_{K-1}} \mathbb{E}^{j_2,\ldots,j_{K-1}}\big[1\big] = \frac{K!}{2}\ ,\end{aligned}$$

where the summations over $j_2, \ldots, j_{K-1}$ are over all possible $(K-2)$–tuples of distinct elements in $\{1, \ldots, K\}$.

Step 6 simply puts all pieces together and lower bounds $\max_\sigma\ \mathbb{E}_\sigma r_n$ by

$$\begin{aligned}&\frac{\mu_1-\mu_2}{K!}\ \sum_\sigma \mathbb{E}_{K,\sigma}\big[1 - p_{\sigma(1),n}\big]\ \mathbb{P}_\sigma\left\{C_{\sigma(1),n} = 0\right\}\ \mathbb{P}_{1,\sigma}\left\{C_{\sigma(K),n} = 0\right\} \\ \geqslant\ &\frac{\mu_1-\mu_2}{2}\ \Big((1-\mu_K)^{C/(\mu_2-\mu_K)}\ (1-\mu_1)^{C/\mu_2}\Big)^{\varepsilon(n)} \ .\end{aligned}$$

Uniform allocation (Unif) — Plays all arms one after the other
For each round $t = 1, 2, \dots,$
pull $I_t = [t \bmod K]$, where $[t \bmod K]$ denotes the value of t modulo K.

UCB — Plays at each round the arm with the highest upper confidence bound
Parameter: exploration factor $\alpha > 1$
For each round $t = 1, 2, \dots,$

(1) for each $i \in \{1, \dots, K\}$, if $T_i(t-1) = 0$ let $B_{i,t} = +\infty$; otherwise, let

$$B_{i,t} = \widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{\alpha \ln t}{T_i(t-1)}} \qquad \text{where} \qquad \widehat{\mu}_{i,T_i(t-1)} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s} \; ;$$

(2) Pull $I_t \in \operatorname*{argmax}_{i=1,\dots,K} B_{i,t}$
(ties broken by choosing, for instance, the arm with smallest index).

Figure 2: Two allocation strategies.

□

4. Upper bounds on the simple regret

In this section, we aim at qualifying the implications of Theorem 6.1 by pointing out that is should be interpreted as a result for large n only. For moderate values of n, strategies not pulling each arm a linear number of times in the exploration phase can have a smaller simple regret.

To do so, we consider only two natural and well-used allocation strategies. The first one is the uniform allocation, which we use as a simple benchmark; it pulls each arm a linear number of times (see Figure 2 for its formal description). The second one is UCB, also described in Figure 2. It is designed for the classical exploration–exploitation dilemma (i.e., it minimizes the cumulative regret) and pulls suboptimal arms a logarithmic number of times only.

In addition to these allocation strategies we consider three recommendation strategies, the ones that recommend respectively the empirical distribution of plays, the empirical best arm, or the most played arm. They are formally defined in Figure 3.

Table 1 summarizes the distribution-dependent and distribution-free bounds we could prove (the difference between the two families of bounds is whether the constants in the bounds can depend or not on the unknown distributions ν_j). In particular, it indicates that while for distribution-dependent bounds, the asymptotic optimal rate of decrease for the simple regret in the number n of rounds is exponential, for distribution-free bounds, this rate worsens to $1/\sqrt{n}$. A similar situation arises for the cumulative regret, see Theorem 2.7 (optimal $\ln n$ rate for distribution-dependent bounds) versus Theorem 2.6 (optimal $\sqrt{n}$ rate for distribution-free bounds).

REMARK 6.1. *The distribution-free lower bound in Table 1 directly follows from the proof of Theorem 2.6. Indeed one can see that the proof goes through for any random variable J_n which is measurable with respect to the history of the forecaster. Thus, one obtains for $n \geqslant K \geqslant 2$,*

$$\inf \sup \mathbb{E} r_n \geqslant \frac{1}{20} \sqrt{\frac{K}{n}} \, ,$$

where the infimum is taken over all forecasters while the supremum considers all sets of K distributions over $[0, 1]$.

Parameters: the history $I_1, \dots, I_n$ of played actions and of their associated rewards $Y_1, \dots, Y_n$, grouped according to the arms as $X_{i,1}, \dots, X_{i,T_i(n)}$, for $i = 1, \dots, n$

Empirical distribution of plays (EDP)
Recommends arm i with probability $T_i(n)/n$, that is, draws J_n at random according to

$$\boldsymbol{p}_n = \left(\frac{T_1(n)}{n}, \dots, \frac{T_K(n)}{n} \right) .$$

Empirical best arm (EBA)
Only considers arms i with $T_i(n) \geqslant 1$, computes their associated empirical means

$$\widehat{\mu}_{i,T_i(n)} = \frac{1}{T_i(n)} \sum_{s=1}^{T_i(n)} X_{i,s} ,$$

and forms the recommendation

$$J_n \in \operatorname*{argmax}_{i=1,\dots,K} \widehat{\mu}_{i,T_i(n)}$$

(ties broken in some way).

Most played arm (MPA)
Recommends the most played arm,

$$J_n \in \operatorname*{argmax}_{i=1,\dots,K} T_i(n)$$

(ties broken in some way).

Figure 3: Three recommendation strategies.

4.1. A simple benchmark: the uniform allocation strategy. As explained above, the combination of the uniform allocation with the recommendation indicating the empirical best arm, forms an important theoretical benchmark. This section studies briefly its theoretical properties: the rate of decrease of its simple regret is exponential in a distribution-dependent sense and equals the optimal (up to a logarithmic term) $1/\sqrt{n}$ rate in the distribution-free case.

Below, we mean by the recommendation given by the empirical best arm at round $K\lfloor n/K \rfloor$ the recommendation $J_{K\lfloor n/K \rfloor}$ of EBA (see Figure 3), where $\lfloor x \rfloor$ denotes the lower integer part of a real number x. The reason why at round n we prefer $J_{K\lfloor n/K \rfloor}$ to J_n is only technical. The analysis is indeed simpler when all averages over the rewards obtained by each arm are over the same number of terms. This happens at rounds n multiple of K and this is why we prefer taking the recommendation of round $K\lfloor n/K \rfloor$ instead of the one of round n.

We propose first two distribution-dependent bounds, the first one is sharper in the case when there are few arms, while the second one is suited for large K.

PROPOSITION 6.1 (Distribution-dependent; Unif and EBA). *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round $K\lfloor n/K \rfloor$) ensures that*

$$\mathbb{E} r_n \leqslant \sum_{i:\Delta_i > 0} \Delta_i \, e^{-\Delta_i^2 \lfloor n/K \rfloor} \qquad \textit{for all } n \geqslant K ;$$

and also, for any $\eta \in (0,1)$,

$$\mathbb{E} r_n \leqslant \left(\max_{i=1,\dots,K} \Delta_i \right) \exp\left(-\frac{(1-\eta)^2}{2} \left\lfloor \frac{n}{K} \right\rfloor \Delta^2 \right) \quad \textit{for all } n \geqslant \left(1 + \frac{2 \ln K}{(1-\eta)^2 \Delta^2} \right) K .$$

	Distribution-dependent		
	EDP	EBA	MPA
Uniform		$\bigcirc e^{-\bigcirc n}$ (Pr.6.1)	
UCB	$\bigcirc(\alpha \ln n)/n$ (Rk.6.2)	$\bigcirc n^{-\bigcirc}$ (Rk.6.3)	$\bigcirc n^{2(1-\alpha)}$ (Th.6.2)
Lower bound		$\bigcirc e^{-\bigcirc n}$ (Cor.6.1)	
	Distribution-free		
	EDP	EBA	MPA
Uniform		$\square\sqrt{\frac{K \ln K}{n}}$ (Cor.6.3)	
UCB	$\square\sqrt{\frac{\alpha K \ln n}{n}}$ (Rk.6.2)	$\frac{\square}{\sqrt{\ln n}}$ (Rk.6.3)	$\square\sqrt{\frac{\alpha K \ln n}{n}}$ (Th.6.3)
Lower bound		$\square\sqrt{\frac{K}{n}}$ (Rk.6.1)	

Table 1: Distribution-dependent (top) and distribution-free (bottom) upper bounds on the expected simple regret of the considered pairs of allocation (rows) and recommendation (columns) strategies. Lower bounds are also indicated. The $\square$ symbols denote the universal constants, whereas the $\bigcirc$ are distribution-dependent constants. In parentheses, we provide the reference (index of the proposition, theorem, remark, corollary) where the stated bound is proved.

PROOF. To prove the first inequality, we relate the simple regret to the probability of choosing a non-optimal arm,

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} = \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{P}\{J_n = i\} \leqslant \sum_{i:\Delta_i>0} \Delta_i \, \mathbb{P}\big\{\widehat{\mu}_{i,\lfloor n/K \rfloor} \geqslant \widehat{\mu}_{i^*,\lfloor n/K \rfloor}\big\}$$

where the upper bound follows from the fact that to be the empirical best arm, an arm i must have performed, in particular, better than a best arm i^*. We now apply Hoeffding's inequality, see Theorem 10.1. We use the fact that we have $2\lfloor n/K \rfloor$ random variables in $[0,1]$. Thus, the probability of interest is bounded by

$$\mathbb{P}\big\{\widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \geqslant 0\big\} \leqslant \exp\left(-\frac{2\,\lfloor n/K \rfloor^2 \,\Delta_i^2}{2\,\lfloor n/K \rfloor}\right),$$

which yields the first result.

The second inequality is proved by resorting to a sharper concentration argument, namely the Mc Diarmid's inequality, see Theorem 10.4. The complete proof can be found in Section 7.3. $\square$

The distribution-free bound of Corollary 6.3 is obtained not directly as a corollary of Proposition 6.1, but as a consequence of its proof. (It is not enough to optimize the bound of Proposition 6.1 over the Δ_i, for it would yield an additional multiplicative factor of K.)

COROLLARY 6.3 (Distribution-free; Unif and EBA). *The uniform allocation strategy associated to the recommendation given by the empirical best arm (at round $K\lfloor n/K \rfloor$) ensures that*

$$\sup_{\nu_1,\dots,\nu_K} \mathbb{E}r_n \leqslant 2\sqrt{\frac{2K \ln K}{n}},$$

where the supremum is over all K-tuples $(\nu_1, \dots, \nu_K)$ of distributions over $[0,1]$.

PROOF. We extract from the proof of Proposition 6.1 that

$$\mathbb{P}\{J_n = i\} \leqslant \exp\left(-\frac{1}{2}\left\lfloor\frac{n}{K}\right\rfloor \Delta_i^2\right) \ ;$$

we now distinguish whether a given Δ_i is more or less than a threshold ε, use that $\sum \mathbb{P}\{J_n = i\} = 1$ and $\Delta_i \leqslant 1$ for all i, to write

$$\begin{aligned}(6.2)\qquad \mathbb{E}r_n = \sum_{i=1}^{K} \Delta_i \, \mathbb{P}\{J_n = i\} &\leqslant \varepsilon + \sum_{i:\Delta_i>\varepsilon} \Delta_i \, \mathbb{P}\{J_n = i\} \\ &\leqslant \varepsilon + \sum_{i:\Delta_i>\varepsilon} \Delta_i \, \exp\left(-\frac{1}{2}\left\lfloor\frac{n}{K}\right\rfloor \Delta_i^2\right) \\ &\leqslant \varepsilon + (K-1)\varepsilon \, \exp\left(-\frac{1}{2}\varepsilon^2 \left\lfloor\frac{n}{K}\right\rfloor\right) \ ,\end{aligned}$$

where the last inequality comes by function study, provided that $\varepsilon \geqslant 1/\lfloor n/K \rfloor$: for $C > 0$, the function $x \in [0,1] \mapsto x \exp(-Cx^2/2)$ is decreasing on $[1/\sqrt{C}, 1]$. Substituting $\varepsilon = \sqrt{(2\ln K)/\lfloor n/K \rfloor}$ concludes the proof. □

4.2. Analysis of UCB as an allocation strategy. We start by studying the recommendation given by the most played arm. A (distribution-dependent) bound is stated in Theorem 6.2; the bound does not involve any quantity depending on the Δ_i, but it only holds for rounds n large enough, a statement that does involve the Δ_i. Its interest is first that it is simple to read, and second, that the techniques used to prove it imply easily a second (distribution-free) bound, stated in Theorem 6.3 and which is comparable to Corollary 6.3.

THEOREM 6.2 (Distribution-dependent; UCB and MPA). *For $\alpha > 1$, the allocation strategy given by UCB associated to the recommendation given by the most played arm ensures that*

$$\mathbb{E}r_n \leqslant \frac{K}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that $n \geqslant K + \dfrac{4K\alpha \ln n}{\Delta^2}$ and $n \geqslant K(K+2)$.

The polynomial rate in the upper bound above is not a coincidence according to the lower bound exhibited in Corollary 6.2. Here, surprisingly enough, this polynomial rate of decrease is distribution-free (but in compensation, the bound is only valid after a distribution-dependent time). This rate illustrates Theorem 6.1: the larger α, the larger the (theoretical bound on the) cumulative regret of UCB but the smaller the simple regret of UCB associated to the recommendation given by the most played arm.

THEOREM 6.3 (Distribution-free; UCB and MPA). *For $\alpha > 1$, the allocation strategy given by UCB associated to the recommendation given by the most played arm ensures that, for all $n \geqslant K(K+2)$,*

$$\sup_{\nu_1,\ldots,\nu_K} \mathbb{E}r_n \leqslant \sqrt{\frac{4K\alpha \ln n}{n-K}} + \frac{K}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)} = O\left(\sqrt{\frac{K\alpha \ln n}{n}}\right) \ ,$$

where the supremum is over all K-tuples $(\nu_1, \ldots, \nu_K)$ of distributions over $[0,1]$.

4.2.1. *Proofs of Theorems 6.2 and 6.3.* We start by a technical lemma from which the two theorems will follow easily.

LEMMA 6.1. *Let $a_1, \dots, a_K$ be real numbers such that $a_1 + \dots + a_K = 1$ and $a_i \geqslant 0$ for all i, with the additional property that for all suboptimal arms i and all optimal arms i^*, one has $a_i \leqslant a_{i^*}$. Then for $\alpha > 1$, the allocation strategy given by UCB associated to the recommendation given by the most played arm ensures that*

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha - 1} \sum_{i \neq i^*} (a_i n - 1)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that, for all suboptimal arms i,

$$a_i n \geqslant 1 + \frac{4\alpha \ln n}{\Delta_i^2} \quad \text{and} \quad a_i n \geqslant K + 2 \, .$$

PROOF. We first prove that whenever the most played arm J_n is different from an optimal arm i^*, then at least one of the suboptimal arms i is such that $T_i(n) \geqslant a_i n$. To do so, we prove the converse and assume that $T_i(n) < a_i n$ for all suboptimal arms. Then,

$$\left(\sum_{i=1}^K a_i \right) n = n = \sum_{i=1}^K T_i(n) < \sum_{i^*} T_{i^*}(n) + \sum_i a_i n$$

where, in the inequality, the first summation is over the optimal arms, the second one, over the suboptimal ones. Therefore, we get

$$\sum_{i^*} a_{i^*} n < \sum_{i^*} T_{i^*}(n)$$

and there exists at least one optimal arm i^* such that $T_{i^*}(n) > a_{i^*} n$. Since by definition of the vector $(a_1, \dots, a_K)$, one has $a_i \leqslant a_{i^*}$ for all suboptimal arms, it comes that $T_i(n) < a_i n \leqslant a_{i^*} n < T_{i^*}(n)$ for all suboptimal arms, and the most played arm J_n is thus an optimal arm.

Thus, using that $\Delta_i \leqslant 1$ for all i,

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} \leqslant \sum_{i : \Delta_i > 0} \mathbb{P}\big\{ T_i(n) \geqslant a_i n \big\} \, .$$

A side-result extracted from the proof of [Audibert et al., 2009, proof of Theorem 7], see also [Auer et al., 2002, proof of Theorem 1], states that for all suboptimal arms i and all rounds $t \geqslant K + 1$,

$$\mathbb{P}\Big\{ I_t = i \ \text{ and } \ T_i(t-1) \geqslant \ell \Big\} \leqslant 2\, t^{1-2\alpha} \qquad \text{whenever} \qquad \ell \geqslant \frac{4\alpha \ln n}{\Delta_i^2} \, . \tag{6.3}$$

Note that this result is weaker than the one proved in Section 2.2 but it is easier to manipulate. We denote by $\lceil x \rceil$ the upper integer part of a real number x. For a suboptimal arm i and since by the assumptions on n and the a_i, the choice $\ell = \lceil a_i n \rceil - 1$ satisfies $\ell \geqslant K + 1$ and $\ell \geqslant (4\alpha \ln n)/\Delta_i^2$,

$$\begin{aligned} \mathbb{P}\big\{ T_i(n) \geqslant a_i n \big\} &= \mathbb{P}\big\{ T_i(n) \geqslant \lceil a_i n \rceil \big\} \\ &\leqslant \sum_{t = \lceil a_i n \rceil}^n \mathbb{P}\Big\{ T_i(t-1) = \lceil a_i n \rceil - 1 \ \text{ and } \ I_t = i \Big\} \\ &\leqslant \sum_{t = \lceil a_i n \rceil}^n 2\, t^{1-2\alpha} \leqslant 2 \int_{\lceil a_i n \rceil - 1}^{\infty} v^{1-2\alpha} \, \mathrm{d}v \leqslant \frac{1}{\alpha - 1} (a_i n - 1)^{2(1-\alpha)} \, , \end{aligned} \tag{6.4}$$

where we used a union bound for the second inequality and (6.3) for the third inequality. A summation over all suboptimal arms i concludes the proof. □

OF THEOREM 6.2. It consists in applying Lemma 6.1 with the uniform choice $a_i = 1/K$ and recalling that Δ is the minimum of the $\Delta_i > 0$. □

OF THEOREM 6.3. We start the proof by using that $\sum \mathbb{P}\{J_n = i\} = 1$ and $\Delta_i \leqslant 1$ for all i, and can thus write

$$\mathbb{E}r_n = \mathbb{E}\Delta_{J_n} = \sum_{i=1}^{K} \Delta_i \, \mathbb{P}\{J_n = i\} \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \Delta_i \, \mathbb{P}\{J_n = i\} \, .$$

Since $J_n = i$ only if $T_i(n) \geqslant n/K$, we get

$$\mathbb{E}r_n \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \Delta_i \, \mathbb{P}\Big\{T_i(n) \geqslant \frac{n}{K}\Big\} \, .$$

Applying (6.4) with $a_i = 1/K$ leads to

$$\mathbb{E}r_n \leqslant \varepsilon + \sum_{i:\Delta_i > \varepsilon} \frac{\Delta_i}{\alpha - 1} \left(\frac{n}{K} - 1\right)^{2(1-\alpha)} \, ,$$

where ε is chosen such that for all $\Delta_i > \varepsilon$, the condition

$$\ell \geqslant n/K - 1 \geqslant (4\alpha \ln n)/\Delta_i^2$$

is satisfied ($n/K - 1 \geqslant K + 1$ being satisfied by the assumption on n and K). The conclusion thus follows from taking, for instance,

$$\varepsilon = \sqrt{(4\alpha K \ln n)/(n - K)}$$

and upper bounding all remaining Δ_i by 1. □

4.2.2. *Other recommendation strategies.* We discuss here the combination of UCB with the two other recommendation strategies, namely, the choice of the empirical best arm and the use of the empirical distribution of plays.

REMARK 6.2 (UCB and EDP). *We indicate in this remark from which results the corresponding bounds of Table 1 follow. As noticed in the beginning of Section 3, in the case of a recommendation formed by the empirical distribution of plays, the simple regret is bounded in terms of the cumulative regret as $\mathbb{E}r_n \leqslant \mathbb{E}R_n/n$. Thus the bounds on $\mathbb{E}r_n$ for UCB and EDP directly follows from Theorem 2.2.*

REMARK 6.3 (UCB and EBA). *We can rephrase the results of Kocsis and Szepesvari [2006] as using UCB as an allocation strategy and forming a recommendation according to the empirical best arm. In particular, [Kocsis and Szepesvari, 2006, Theorem 5] provides a distribution-dependent bound on the probability of not picking the best arm with this procedure and can be used to derive the following bound on the simple regret of UCB combined with EBA: for all $n \geqslant 1$,*

$$\mathbb{E}r_n \leqslant \sum_{i:\Delta_i > 0} \frac{4}{\Delta_i} \left(\frac{1}{n}\right)^{\rho_\alpha \Delta_i^2/2}$$

where ρ_α is a positive constant depending on α only. The leading constants $1/\Delta_i$ and the distribution-dependent exponent make it not as useful as the one presented in Theorem 6.2. The best distribution-free bound we could get from this bound was of the order of $1/\sqrt{\rho_\alpha \ln n}$, to be compared to the asymptotic optimal $1/\sqrt{n}$ rate stated in Theorem 6.3.

5. Conclusions: Comparison of the bounds, simulation study

We first explain why, in some cases, the bound provided by our theoretical analysis in Lemma 6.1 (for UCB and MPA) is better than the bound stated in Proposition 6.1 (for Unif and EBA). The central point in the argument is that the bound of Lemma 6.1 is of the form $\bigcirc\, n^{2(1-\alpha)}$, for some

distribution-dependent constant $\bigcirc$, that is, it has a distribution-free convergence rate. In comparison, the bound of Proposition 6.1 involves the gaps Δ_i in the rate of convergence. Some care is needed in the comparison, since the bound for UCB holds only for n large enough, but it is easy to find situations where for moderate values of n, the bound exhibited for the sampling with UCB is better than the one for the uniform allocation. These situations typically involve a rather large number K of arms; in the latter case, the uniform allocation strategy only samples $\lfloor n/K \rfloor$ times each arm, whereas the UCB strategy focuses rapidly its exploration on the best arms. A general argument is proposed in Section 7.4 as well as a numerical example, showing that for moderate values of n, the bounds associated to the sampling with UCB are better than the ones associated to the uniform sampling. This is further illustrated numerically, in the right part of Figure 4).

To make short the longer story described in this chapter, one can distinguish three regimes, according to the value of the number of rounds n. The statements of these regimes (the ranges of their corresponding n) involve distribution-dependent quantifications, to determine which n are considered small, moderate, or large.

- For large values of n, uniform exploration is better (as shown by a combination of the lower bound of Corollary 6.2 and of the upper bound of Proposition 6.1).
- For moderate values of n, sampling with UCB is preferable, as discussed just above (and in Section 7.4).
- For small values of n, little can be said and the best bounds to consider are perhaps the distribution-free bounds, which are of the same order of magnitude for the two pairs of strategies.

We propose two simple experiments to illustrate our theoretical analysis; each of them was run on 10^4 instances of the problem and we plotted the average simple regret. This is an instance of the Monte-Carlo method and provides accurate estimators of the expected simple regret $\mathbb{E}r_n$.

The first experiment (upper plot of Figure 4) shows that for small values of n (here, $n \leqslant 80$), the uniform allocation strategy can have an interesting behavior. Of course the range of these "small" values of n can be made arbitrarily large by decreasing the gap Δ. The second one (lower plot of Figure 4) corresponds to the numerical example to be described in Section 7.4. In both cases, the unclear picture for small values of n become clearer for moderate values and shows an advantage in favor of UCB–based strategies.

REMARK 6.4. *We mostly illustrated here the small and moderate n regimes. This is because for large n, the simple regret is usually very small, even below computer precision. Therefore, because of the chosen ranges, we do not see yet the uniform allocation strategy getting better than UCB–based strategies, a fact that is true however for large enough n. This has an important impact on the interpretation of the lower bound of Theorem 6.1. While its statement is in finite time, it should be interpreted as providing an asymptotic result only.*

6. Pure exploration for $\mathcal{X}$–armed bandits

This section is of theoretical interest. We consider the $\mathcal{X}$-armed bandit problem, of Chapter 4 and (re)define the notions of cumulative and simple regret in this setting. We show that the cumulative regret can be minimized if and only if the simple regret can be minimized, and use this equivalence to characterize the metric spaces $\mathcal{X}$ in which the cumulative regret can be minimized: the separable ones. Here, in addition to its natural interpretation, the simple regret thus appears as a tool for proving results on the cumulative regret.

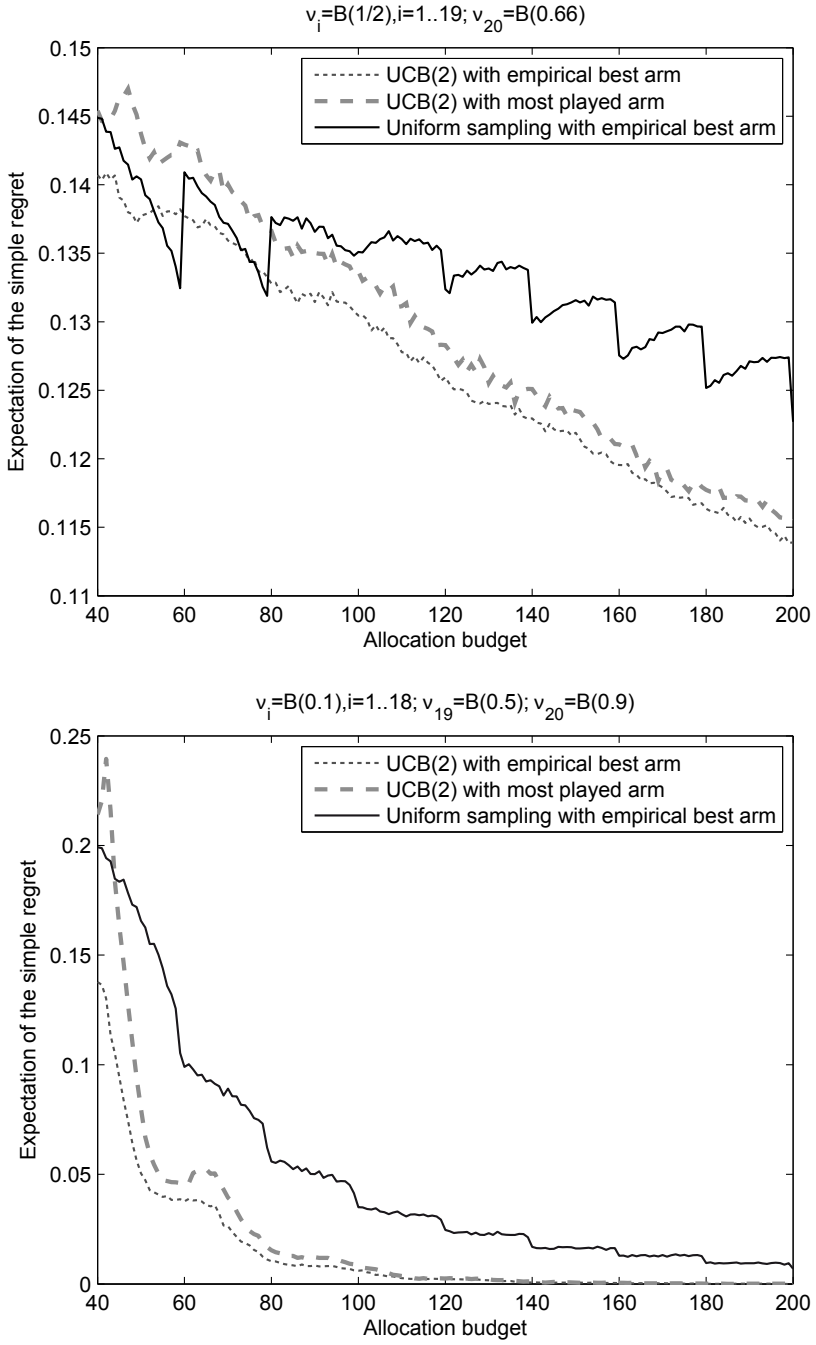

Figure 4: $K = 20$ arms with Bernoulli distributions of parameters indicated on top of each graph. x-axis: number of rounds n; y-axis: simple regrets $\mathbb{E}r_n$ (estimated by a Monte-Carlo method).

6.1. Description of the model of $\mathcal{X}$–armed bandits. We consider a bounded interval of $\mathbb{R}$, say $[0, 1]$ again. We denote by $\mathcal{P}([0, 1])$ the set of probability distributions over $[0, 1]$. Similarly, given a topological space $\mathcal{X}$, we denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$. We then call environment on $\mathcal{X}$ any mapping $E : \mathcal{X} \to \mathcal{P}([0, 1])$. We say that E is continuous if the mapping that associates to each $x \in \mathcal{X}$ the expectation $\mu(x)$ of $E(x)$ is continuous.

Parameter: an unknown environment $E : \mathcal{X} \to \mathcal{P}([0,1])$

For each round $t = 1, 2, \ldots,$

(1) the forecaster chooses a distribution $\varphi_t \in \mathcal{P}(\mathcal{X})$ and pulls an arm I_t at random according to φ_t;

(2) the environment draws the reward Y_t for that action, according to $E(I_t)$.

Goal:

Find an allocation strategy (φ_t) such that the cumulative regret

$$R_n = n \sup_{x \in \mathcal{X}} \mu(x) - \sum_{t=1}^{n} \mu(I_t)$$

is small (i.e., $o(n)$, in expectation).

Figure 5: The anytime $\mathcal{X}$–armed bandit problem.

Parameter: an unknown environment $E : \mathcal{X} \to \mathcal{P}([0,1])$

For each round $t = 1, 2, \ldots,$

(1) the forecaster chooses a distribution $\varphi_t \in \mathcal{P}(\mathcal{X})$ and pulls an arm I_t at random according to φ_t;

(2) the environment draws the reward Y_t for that action, according to $E(I_t)$;

(3) the forecaster outputs a recommendation $\psi_t \in \mathcal{P}(\mathcal{X})$;

(4) if the environment sends a stopping signal, then the game takes an end; otherwise, the next round starts.

Goal:

Find an allocation strategy (φ_t) and a recommendation strategy (ψ_t) such that the simple regret

$$r_n = \sup_{x \in \mathcal{X}} \mu(x) - \int_{\mathcal{X}} \mu(x) \, \mathrm{d}\psi_n(x)$$

is small (i.e., $o(1)$, in expectation).

Figure 6: The anytime pure exploration problem for $\mathcal{X}$–armed bandits.

The $\mathcal{X}$–armed bandit problem is described in Figures 5 and 6. There, an environment E on $\mathcal{X}$ is fixed and we want various notions of regret to be small, given this environment.

We consider now families of environments and say that a family $\mathcal{F}$ of environments is explorable–exploitable (respectively, explorable) if there exists a forecaster such that for any environment $E \in \mathcal{F}$, the expected cumulative regret $\mathbb{E}R_n$ (expectation taken with respect to E and all auxiliary randomizations) is $o(n)$ (respectively, $\mathbb{E}r_n = o(1)$). Of course, explorability of $\mathcal{F}$ is a milder requirement than explorability–exploitability of $\mathcal{F}$, as can be seen by considering the recommendation given by the empirical distribution of plays of Figure 3 and applying the same argument as the one used at the beginning of Section 3.

In fact, it can be seen that the two notions are equivalent, and this is why we will henceforth concentrate on explorability only, for which characterizations as the ones of Theorem 6.4 are simpler to exhibit and prove.

LEMMA 6.2. *A family of environments $\mathcal{F}$ is explorable if and only if it is explorable–exploitable.*

The proof can be found in Section 7.1. It relies essentially on designing a strategy suited for cumulative regret from a strategy minimizing the simple regret; to do so, exploration and exploitation occur at fixed rounds in two distinct phases and only the payoffs obtained during exploration rounds are fed into the base allocation strategy.

6.2. A positive result for metric spaces. We denote by $\mathcal{P}([0,1])^{\mathcal{X}}$ the family of all possible environments E on $\mathcal{X}$, and by $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ the subset of $\mathcal{P}([0,1])^{\mathcal{X}}$ formed by the continuous environments.

EXAMPLE 6.1. *Previous sections were about the family $\mathcal{P}([0,1])^{\mathcal{X}}$ of all environments over $\mathcal{X} = \{1, \dots, K\}$ being explorable.*

The main result concerning $\mathcal{X}$–armed bandit problems is formed by the following equivalences in metric spaces. It generalizes the result of Example 6.1.

THEOREM 6.4. *Let $\mathcal{X}$ be a metric space. Then $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ is explorable if and only if $\mathcal{X}$ is separable.*

COROLLARY 6.4. *Let $\mathcal{X}$ be a set. $\mathcal{P}([0,1])^{\mathcal{X}}$ is explorable if and only if $\mathcal{X}$ is countable.*

The proofs can be found in Section 7.2. Their main technical ingredient is that there exists a probability distribution over a metric space $\mathcal{X}$ giving a positive probability mass to all open sets if and only if $\mathcal{X}$ is separable. Then, whenever it exists, it allows some uniform exploration.

7. Technical Proofs

7.1. Proof of Lemma 6.2.

PROOF. In view of the comments before the statement of Lemma 6.2, we need only to prove that an explorable family $\mathcal{F}$ is also explorable–exploitable. We consider a pair of allocation (φ_t) and recommendation (ψ_t) strategies such that for all environments $E \in \mathcal{F}$, the simple regret satisfy $\mathbb{E}r_n = o(1)$, and provide a new strategy (φ'_t) such that its cumulative regret satisfies $\mathbb{E}R'_n = o(n)$ for all environments $E \in \mathcal{F}$.

It is defined informally as follows. At round $t = 1$, it uses $\varphi'_1 = \varphi_1$ and gets a reward Y_1. Based on this reward, the recommendation $\psi_1(Y_1)$ is formed and at round $t = 2$, the new strategy plays $\varphi'_2(Y_1) = \psi_1(Y_1)$. It gets a reward Y_2 but does not take it into account. It bases its choice $\varphi'_3(Y_1, Y_2) = \varphi_2(Y_1)$ only on Y_1 and gets a reward Y_3. Based on Y_1 and Y_3, the recommendation $\psi_2(Y_1, Y_3)$ is formed and played at rounds $t = 4$ and $t = 5$, i.e.,

$$\varphi'_4(Y_1, Y_2, Y_3) = \varphi'_5(Y_1, Y_2, Y_3, Y_4) = \psi_2(Y_1, Y_3)\,.$$

And so on: the sequence of distributions chosen by the new strategy is formed using the applications

$$\begin{aligned}
&\varphi_1, \quad \psi_1,\\
&\varphi_2, \quad \psi_2,\ \psi_2,\\
&\varphi_3, \quad \psi_3,\ \psi_3,\ \psi_3,\\
&\varphi_4, \quad \psi_4,\ \psi_4,\ \psi_4,\ \psi_4,
\end{aligned}$$

$$\varphi_5, \quad \psi_5, \psi_5, \psi_5, \psi_5, \psi_5,$$
$$\dots$$

Formally, we consider regimes indexed by integers $t \geqslant 1$ and of length $1+t$. The t–th regime starts at round

$$1 + \sum_{s=1}^{t-1}(1+s) = t + \frac{t(t-1)}{2} = \frac{t(t+1)}{2} \, .$$

During this regime, the following distributions are used,

$$\varphi'_{t(t+1)/2+k} = \begin{cases} \varphi_t\Big(\big(Y_{s(s+1)/2}\big)_{s=1,\dots,t-1}\Big) & \text{if } k = 0; \\ \psi_t\Big(\big(Y_{s(s+1)/2}\big)_{s=1,\dots,t-1}\Big) & \text{if } 1 \leqslant k \leqslant t. \end{cases}$$

Note that we only keep track of the payoffs obtained when $k = 0$ in a regime.

The regret R'_n at round n of this strategy is as follows. We decompose n in a unique manner as

$$n = \frac{t(n)\big(t(n)+1\big)}{2} + k(n) \qquad \text{where} \qquad k(n) \in \big\{0, \dots, t(n)\big\} \, . \tag{6.5}$$

Then (using also the tower rule),

$$\mathbb{E}R'_n \leqslant t(n) + \Big(\mathbb{E}r_1 + 2\,\mathbb{E}r_2 + \dots + \big(t(n)-1\big)\,\mathbb{E}r_{t(n)-1} + k(n)\,\mathbb{E}r_{t(n)}\Big)$$

where the first term comes from the time rounds when the new strategy used the base allocation strategy to explore and where the other terms come from the ones when it exploited. This inequality can be rewritten as

$$\frac{\mathbb{E}R'_n}{n} \leqslant \frac{t(n)}{n} + \frac{k(n)\,\mathbb{E}r_{t(n)} + \sum_{s=1}^{t(n)-1} s\,\mathbb{E}r_s}{n} \, ,$$

which shows that $\mathbb{E}R'_n = o(n)$ whenever $\mathbb{E}r_s = o(1)$ as $s \to \infty$, since the first term in the right-hand side is of the order of $1/\sqrt{n}$ and the second one is a Cesaro average. This concludes that the exhibited strategy has a small cumulative regret for all environments of the family, which is thus explorable–exploitable. □

7.2. Proof of Theorem 6.4 and its corollary. The key ingredient is the following characterization of separability (which relies on an application of Zorn's lemma); see, e.g., [Billingsley, 1968, Appendix I, page 216].

LEMMA 6.3. *Let $\mathcal{X}$ be a metric space, with distance denoted by d. $\mathcal{X}$ is separable if and only if it contains no uncountable subset A such that*

$$\rho = \inf\big\{d(x,y) : x, y \in A\big\} > 0 \, .$$

Separability can then be characterized in terms of the existence of a probability distribution with full support. Though it seems natural, we did not see any reference to it in the literature and this is why we state it. (In the proof of Theorem 6.4, we will only use the straightforward direct part of the characterization.)

LEMMA 6.4. *Let $\mathcal{X}$ be a metric space. There exists a probability distribution λ on $\mathcal{X}$ with $\lambda(V) > 0$ for all open sets V if and only if $\mathcal{X}$ is separable.*

PROOF. We prove the converse implication first. If $\mathcal{X}$ is separable, we denote by $x_1, x_2, \dots$ a dense sequence. If it is finite with length N, we let

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$$

and otherwise,

$$\lambda = \sum_{i \geqslant 1} \frac{1}{2^i} \delta_{x_i} \, .$$

The result follows, since each open set V contains at least some x_i.

For the direct implication, we use Lemma 6.3 (and its notations). If $\mathcal{X}$ is not separable, then it contains uncountably many disjoint open balls, formed by the $B(a, \rho/2)$, for $a \in A$. If there existed a probability distribution λ with full support on $\mathcal{X}$, it would in particular give a positive probability to all these balls; but this is impossible, since there are uncountably many of them. □

7.2.1. *Separability of $\mathcal{X}$ implies explorability of the family $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$.* The proof of the converse part of the characterization provided by Theorem 6.4 relies on a somewhat uniform exploration. We reach each open set of $\mathcal{X}$ in a geometric time.

PROOF. Since $\mathcal{X}$ is separable, there exists a probability distribution λ on $\mathcal{X}$ with $\lambda(V) > 0$ for all open sets V, as asserted by Lemma 6.4.

The proposed strategy is then constructed in a way similar to the one exhibited in Section 7.4, in the sense that we also consider successives regimes, where the t–th of them has also length $1+t$. They use the following allocations,

$$\varphi_{t(t+1)/2+k} = \begin{cases} \lambda & \text{if } k = 0; \\ \delta_{I_{k(k+1)/2}} & \text{if } 1 \leqslant k \leqslant t. \end{cases}$$

Put in words, at the beginning of each regime, a new point $I_{t(t+1)/2}$ is drawn at random in $\mathcal{X}$ according to λ, and then, all previously drawn points $I_{s(s+1)/2}$, for $1 \leqslant s \leqslant t-1$, and the new point $I_{t(t+1)/2}$ are pulled again, one after the other.

The recommendations ψ_n are deterministic and put all probability mass on the best empirical arm among the first played $g(n)$ arms (where the function g will be determined by the analysis). Formally, for all $x \in \mathcal{X}$ such that

$$T_n(x) = \sum_{t=1}^{n} \mathbb{I}_{\{I_t = x\}} \geqslant 1 \, ,$$

one defines

$$\widehat{\mu}_n(x) = \frac{1}{T_n(x)} \sum_{t=1}^{n} Y_t \, \mathbb{I}_{\{I_t = x\}} \, .$$

Then,

$$\psi_n = \delta_{X_n^*} \quad \text{where} \quad X_n^* \in \operatorname*{argmax}_{1 \leqslant s \leqslant g(n)} \widehat{\mu}_n\big(I_{s(s+1)/2}\big)$$

(ties broken in some way, as usual; and $g(n)$ to be chosen small enough so that all considered arms have been played at least once). Note that exploration and exploitation appear in two distinct phases, as was the case already, for instance, in Section 4.1.

We now denote

$$\mu^* = \sup_{x \in \mathcal{X}} \mu(x) \quad \text{and} \quad \mu^*_{g(n)} = \max_{1 \leqslant s \leqslant g(n)} \mu\big(I_{s(s+1)/2}\big) \, ;$$

the simple regret can then be decomposed as

$$\mathbb{E}r_n = \mu^* - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big] = \Big(\mu^* - \mathbb{E}\Big[\mu_{g(n)}^*\Big]\Big) + \Big(\mathbb{E}\Big[\mu_{g(n)}^*\Big] - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big]\Big) \,,$$

where the first difference can be thought of as an approximation error, and the second one, as resulting from an estimation error. We now show that both differences vanish in the limit.

We first deal with the approximation error. We fix $\varepsilon > 0$. Since μ is continuous on $\mathcal{X}$, there exists an open set V such that

$$\forall x \in V, \qquad \mu^* - \mu(x) \leqslant \varepsilon \,.$$

It follows that

$$\mathbb{P}\Big\{\mu^* - \mu_{g(n)}^* > \varepsilon\Big\} \leqslant \mathbb{P}\Big\{\forall\, s \in \big\{1, \dots, g(n)\big\}, \quad I_{s(s+1)/2} \notin V\Big\}$$
$$\leqslant \big(1 - \lambda(V)\big)^{g(n)} \longrightarrow 0$$

provided that $g(n) \to \infty$ (a condition that will be satisfied, see below). Since in addition, $\mu_{g(n)}^* \leqslant \mu^*$, we get

$$\limsup\ \mu^* - \mathbb{E}\Big[\mu_{g(n)}^*\Big] \leqslant \varepsilon \,.$$

For the difference resulting from the estimation error, we denote

$$I_n^* \in \operatorname*{argmax}_{1 \leqslant s \leqslant g(n)} \mu\big(I_{s(s+1)/2}\big)$$

(ties broken in some way). Fix an arbitrary $\varepsilon > 0$. We note that if for all $1 \leqslant s \leqslant g(n)$,

$$\Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| \leqslant \varepsilon \,,$$

then (together with the definition of X_n^*)

$$\mu\big(X_n^*\big) \geqslant \widehat{\mu}_n\big(X_n^*\big) - \varepsilon \geqslant \widehat{\mu}_n\big(I_n^*\big) - \varepsilon \geqslant \mu\big(I_n^*\big) - 2\varepsilon \,.$$

Thus, we have proved the inequality

$$\mathbb{E}\Big[\mu_{g(n)}^*\Big] - \mathbb{E}\Big[\mu\big(X_n^*\big)\Big] \leqslant 2\varepsilon + \mathbb{P}\Big\{\exists\, s \leqslant g(n), \Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon\Big\} \,. \tag{6.6}$$

We use a union bound and control each (conditional) probability

$$\mathbb{P}\Big\{\Big|\widehat{\mu}_n\big(I_{s(s+1)/2}\big) - \mu\big(I_{s(s+1)/2}\big)\Big| > \varepsilon \ \Big|\ \mathcal{A}_n\Big\} \tag{6.7}$$

for $1 \leqslant s \leqslant g(n)$, where $\mathcal{A}_n$ is the σ–algebra generated by the randomly drawn points $I_{k(k+1)/2}$, for those k with $k(k+1)/2 \leqslant n$. Conditionnally to them, $\widehat{\mu}_n\big(I_{s(s+1)/2}\big)$ is an average of a deterministic number of summands, which only depends on s, and thus, classical concentration-of-the-measure arguments can be used. For instance, the quantities (6.7) are bounded, via an application of Hoeffding's inequality, see Theorem 10.1, by

$$2\exp\Big(-2\,T_n\big(I_{s(s+1)/2}\big)\,\varepsilon^2\Big) \,.$$

We lower bound $T_n\big(I_{s(s+1)/2}\big)$. The point $I_{s(s+1)/2}$ was pulled twice in regime s, once in each regime $s+1, \dots, t(n)-1$, and maybe in $t(n)$, where n is decomposed again as in (6.5). That is,

$$T_n\big(I_{s(s+1)/2}\big) \geqslant t(n) - s + 1 \geqslant \sqrt{2n} - 1 - g(n) \,,$$

since we only consider $s \leqslant g(n)$ and since (6.5) implies that

$$n \leqslant \frac{t(n)\,\big(t(n)+3\big)}{2} \leqslant \frac{\big(t(n)+2\big)^2}{2} \,, \quad \text{that is,}\ \ t(n) \geqslant \sqrt{2n} - 2 \,.$$

Substituting this in the Hoeffding's bound, integrating, and taking a union bound lead from (6.6) to

$$\mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 2\varepsilon + 2g(n)\,\exp\Big(-2\left(\sqrt{2n}-1-g(n)\right)\varepsilon^2\Big)\ .$$

Choosing for instance $g(n) = \sqrt{n}/2$ ensures that

$$\limsup\ \mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 2\varepsilon\ .$$

Summing up the two superior limits, we finally get

$$\limsup\ \mathbb{E}r_n \leqslant \limsup\ \mu^* - \mathbb{E}\Big[\mu^*_{g(n)}\Big] + \limsup\ \mathbb{E}\Big[\mu^*_{g(n)}\Big] - \mathbb{E}\Big[\mu\big(X^*_n\big)\Big] \leqslant 3\varepsilon\ ;$$

since this is true for all arbitrary $\varepsilon > 0$, the proof is concluded. □

7.2.2. *Separability of $\mathcal{X}$ is a necessary condition.* We now prove the direct part of the characterization provided by Theorem 6.4. It basically follows from the impossibility of a uniform exploration, as asserted by Lemma 6.4.

PROOF. Let $\mathcal{X}$ be a non-separable metric space (with distance denoted by d). Let A be an uncountable set and let $\rho > 0$ be defined as in Lemma 6.3; in particular, the balls $B(a, \rho/2)$ are disjoint, for $a \in A$.

We now consider the subset of $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ formed by the environments E_a defined as follows. They are indexed by $a \in A$ and their corresponding expectations are given by

$$\mu_a : x \in \mathcal{X} \longmapsto \left(1 - \frac{d(x,a)}{\rho/2}\right)^+ .$$

Note that μ_a is continuous, that $\mu_a(x) > 0$ for all $x \in B(a, \rho/2)$ but $\mu_a(x) = 0$ for all $x \in \mathcal{X} \setminus B(a, \rho/2)$; that the best arm is a and gets a reward $\mu^*_a = \mu_a(a) = 1$. The associated environment E_a is deterministic, in the sense that it is defined as $E_a(x) = \delta_{\mu_a(x)}$.

We fix a forecaster and denote by $\mathbb{E}_a$ the expectation under environment E_a with respect with the auxiliary randomizations used by the forecaster. By construction of μ_a,

$$\mathbb{E}_a r_n = 1 - \mathbb{E}_a\left[\int_{\mathcal{X}} \mu_a(x)\,\mathrm{d}\psi_n(x)\right] \geqslant 1 - \mathbb{E}_a\Big[\psi_n\big(B(a,\rho/2)\big)\Big]\ .$$

We now show the existence of a non-empty set A' such that for all $a \in A'$ and $n \geqslant 1$,

$$\mathbb{E}_a\Big[\psi_n\big(B(a,\rho/2)\big)\Big] = 0\ ; \tag{6.8}$$

this indicates that $\mathbb{E}_a r_n = 1$ for all $n \geqslant 1$ and $a \in A'$, thus preventing in particular $\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)$ from being explorable by the fixed forecaster.

The set A' is constructed by studying the behavior of the forecaster under the environment E_0 yielding deterministic null rewards throughout the space, i.e., associated to the expectations $x \in \mathcal{X} \mapsto \mu_0(x) = 0$. In the first round, the forecaster chooses a deterministic distribution $\varphi_1 = \varphi^0_1$ over $\mathcal{X}$, picks I_1 at random according to φ^0_1, gets a deterministic payoff $Y_1 = 0$, and finally recommends $\psi^0_1(I_1) = \psi_1(I_1, Y_1)$ (which depends on I_1 only, since the obtained payoffs are all null). In the second round, it chooses an allocation $\psi^0_2(I_1)$ (that depends only on I_1, for the same reasons as before), picks I_2 at random according to $\psi^0_2(I_1)$, gets a null reward, and recommends $\psi^0_2(I_1, I_2)$; and so on.

We denote by $\mathbb{A}$ the probability distribution giving the auxiliary randomizations used to draw the I_t at random, and for all integers t and all measurable applications

$$\nu : (x_1, \ldots, x_t) \in \mathcal{X}^t \longmapsto \nu(x_1, \ldots, x_t) \in \mathcal{P}(\mathcal{X})$$

we introduce the distributions $\mathbb{A}\cdot\nu\in\mathcal{P}(\mathcal{X})$ defined as follows. For all measurable sets $V\subseteq\mathcal{X}$,

$$\mathbb{A}\cdot\nu(V)=\mathbb{E}_{\mathbb{A}}\left[\int_{\mathcal{X}}\mathbb{I}_V\,\mathrm{d}\nu(I_1,\dots,I_t)\right]\,.$$

Now, let B_n and C_n be defined as the at most countable sets of a such that, respectively, $\mathbb{A}\cdot\varphi_n^0$ and $\mathbb{A}\cdot\psi_n^0$ give a positive probability mass to $B(a,\rho/2)$; we recall that the latter is the support of the expectation mapping μ_a. Then, let

$$A'=A\setminus\left(\bigcup_{n\geqslant1}B_n\cup\bigcup_{n\geqslant1}C_n\right)$$

be the uncountable, thus non empty, set of those elements of A which are in no B_n or C_n.

By construction, for all $a\in A'$, the forecaster then behaves similarly under the environments E_a and E_0, since it only gets null rewards (a is in no B_n); this similar behavior means formally that for all measurable sets $V\subseteq\mathcal{X}$ and all $n\geqslant1$,

$$\mathbb{E}_a\big[\varphi_n(V)\big]=\mathbb{A}\cdot\varphi_n^0(V)\quad\text{and}\quad\mathbb{E}_a\big[\psi_n(V)\big]=\mathbb{A}\cdot\psi_n^0(V)\,.$$

In particular, since a is in no C_n, it hits in no recommendation ψ_n the ball $B(a,\rho/2)$, which is exactly what remained to be proved, see (6.8). □

7.2.3. *The countable case of Corollary 6.4.* We adopt an "à la Bourbaki" approach and derive this special case from the general theory.

PROOF. We endow $\mathcal{X}$ with the discrete topology, i.e., choose the distance

$$d(x,y)=\mathbb{I}_{\{x\neq y\}}\,.$$

Then, all applications defined on $\mathcal{X}$ are continuous; in particular,

$$\mathcal{C}\big(\mathcal{P}([0,1])^{\mathcal{X}}\big)=\mathcal{P}([0,1])^{\mathcal{X}}\,.$$

In addition, $\mathcal{X}$ is then separable if and only if it is countable. The result thus follows immediately from Theorem 6.4. □

7.2.4. *An additional remark.* In this chapter, we mostly consider non-uniform bounds (bounds that are individual as far as the environments are concerned). As for uniform bounds, i.e., bounds on quantities of the form

$$\sup_{E\in\mathcal{F}}\mathbb{E}R_n\qquad\text{or}\qquad\sup_{E\in\mathcal{F}}\mathbb{E}r_n$$

for some family $\mathcal{F}$, two observations can be made.

First, it is easy to see that no sublinear uniform bound can be obtained for the family of all continuous environments, as soon as there exists infinitely many disjoint open balls.

However one can exhibit such sublinear uniform bounds in some specific scenarios; for instance, when $\mathcal{X}$ is totally bounded and $\mathcal{F}$ is formed by continuous functions with a common bounded Lipschitz constant.

7.3. Proof of the second statement of Proposition 6.1. We use below the notations introduced in the proof of the first statement of Proposition 6.1.

PROOF. Since some regret is suffered only when an arm with suboptimal expectation has the best empirical performance,

$$\mathbb{E}r_n\leqslant\left(\max_{i=1,\dots,K}\Delta_i\right)\mathbb{P}\left\{\max_{i:\Delta_i>0}\widehat{\mu}_{i,\lfloor n/K\rfloor}\geqslant\widehat{\mu}_{i^*,\lfloor n/K\rfloor}\right\}\,.$$

Now, the quantity of interest can be rewritten as

$$\left\lfloor \frac{n}{K} \right\rfloor \left(\max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right) = f\left(\vec{X}_1, \ldots, \vec{X}_{\lfloor \frac{n}{K} \rfloor} \right)$$

for some function f, where for all $s = 1, \ldots, \lfloor n/K \rfloor$, we denote by $\vec{X}_s$ the vector $(X_{1,s}, \ldots, X_{K,s})$. ($f$ is defined as a maximum of at most $K-1$ sums of differences.) We apply the method of bounded differences, see 10.4. It is straightforward that, since all random variables of interest take values in $[0,1]$, the bounded differences condition is satisfied with ranges all equal to 2. Therefore, the indicated concentration inequality states that

$$\mathbb{P}\left\{ \left(\max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right) - \mathbb{E}\left[\max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right] \geqslant \varepsilon \right\} \leqslant \exp\left(-\frac{2\lfloor n/K \rfloor \varepsilon^2}{4} \right)$$

for all $\varepsilon > 0$. We choose

$$\varepsilon = -\mathbb{E}\left[\max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right] \geqslant \min_{i:\Delta_i>0} \Delta_i - \mathbb{E}\left[\max_{i:\Delta_i>0} \left\{ \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} + \Delta_i \right\} \right]$$

(where we used that the maximum of K first quantities plus the minimum of K other quantities is less than the maximum of the K sums). We now argue that

$$\mathbb{E}\left[\max_{i:\Delta_i>0} \left\{ \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} + \Delta_i \right\} \right] \leqslant \sqrt{\frac{2\ln K}{\lfloor n/K \rfloor}} \, ;$$

this is done by a classical argument, using bounds on the moment generating function of the random variables of interest. Consider

$$Z_i = \lfloor n/K \rfloor \left(\widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} + \Delta_i \right)$$

for all $i = 1, \ldots, K$. Independence and Hoeffding's lemma (see Lemma 10.1) imply that for all $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda Z_i} \right] \leqslant \exp\left(-\frac{1}{2}\lambda^2 \lfloor n/K \rfloor \right)$$

(where we used again that Z_i is given by a sum of random variables bounded between -1 and 1). A well-known inequality for maxima of subgaussian random variables (see, again, [Devroye and Lugosi, 2001, Chapter 2]) then yields

$$\mathbb{E}\left[\max_{i=1,\ldots,K} Z_i \right] \leqslant \sqrt{2\lfloor n/K \rfloor \ln K} \, ,$$

which leads to the claimed upper bound. Putting things together, we get that for the choice

$$\varepsilon = -\mathbb{E}\left[\max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} - \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right] \geqslant \min_{i:\Delta_i>0} \Delta_i - \sqrt{\frac{2\ln K}{\lfloor n/K \rfloor}} > 0$$

(for n sufficiently large, a statement made precise below), we have

$$\begin{aligned} \mathbb{P}\left\{ \max_{i:\Delta_i>0} \widehat{\mu}_{i,\lfloor n/K \rfloor} \geqslant \widehat{\mu}_{i^*,\lfloor n/K \rfloor} \right\} &\leqslant \exp\left(-\frac{2\lfloor n/K \rfloor \varepsilon^2}{4} \right) \\ &\leqslant \exp\left(-\frac{1}{2}\left\lfloor \frac{n}{K} \right\rfloor \left(\min_{i:\Delta_i>0} \Delta_i - \sqrt{\frac{2\ln K}{\lfloor n/K \rfloor}} \right)^2 \right) . \end{aligned}$$

The result follows for n such that

$$\min_{i:\Delta_i>0} \Delta_i - \sqrt{\frac{2\ln K}{\lfloor n/K \rfloor}} \geqslant (1-\eta) \min_{i:\Delta_i>0} \Delta_i \, ;$$

the second part of the theorem indeed only considers such n. □

7.4. Detailed discussion of the heuristic arguments presented in Section 5. We first state the following corollary to Lemma 6.1.

THEOREM 6.5. *The allocation strategy given by UCB (with $\alpha > 1$) associated to the recommendation given by the most played arm ensures that*

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha - 1} \sum_{i \neq i^*} \left(\frac{\beta n}{\Delta_i^2} - 1 \right)^{2(1-\alpha)}$$

for all n sufficiently large, e.g., such that

$$\frac{n}{\ln n} \geqslant \frac{4\alpha + 1}{\beta} \quad \textit{and} \quad n \geqslant \frac{K+2}{\beta} (\Delta')^2 \,,$$

where $\Delta' = \max_i \Delta_i$ and we denote by K^* the number of optimal arms and

$$\beta = \frac{1}{\dfrac{K^*}{\Delta^2} + \displaystyle\sum_{i \neq i^*} \frac{1}{\Delta_i^2}} \,.$$

PROOF. We apply Lemma 6.1 with the choice $a_i = \beta/\Delta_i^2$ for all suboptimal arms i and $a_{i^*} = \beta/\Delta^2$ for all optimal arms i^*, where β denotes the renormalization constant. □

For illustration, consider the case when there is one optimal arm, one Δ–suboptimal arm and $K - 2$ arms that are 2Δ–suboptimal. Then

$$\frac{1}{\beta} = \frac{2}{\Delta^2} + \frac{K-2}{(2\Delta)^2} = \frac{6+K}{4\Delta^2} \,,$$

and the previous bound of Theorem 6.5 implies that

$$\mathbb{E}r_n \leqslant \frac{1}{\alpha - 1} \left(\frac{4n}{6+K} - 1 \right)^{2(1-\alpha)} + \frac{K-2}{\alpha - 1} \left(\frac{n}{6+K} - 1 \right)^{2(1-\alpha)} \tag{6.9}$$

for all n sufficiently large, e.g.,

$$n \geqslant \max \left\{ (K+2)(6+K),\ (4\alpha+1) \left(\frac{6+K}{4\Delta^2} \right) \ln n \right\} \,. \tag{6.10}$$

Now, the upper bound on $\mathbb{E}r_n$ given in Proposition 6.1 for the uniform allocation associated to the recommendation provided by the empirical best arm is larger than

$$\Delta e^{-\Delta^2 \lfloor n/K \rfloor / 2} \,, \qquad \text{for all } n \geqslant K.$$

Thus for n moderately large, e.g., such that $n \geqslant K$ and

$$\lfloor n/K \rfloor \leqslant (4\alpha + 1) \left(\frac{6+K}{4\Delta^2} \right) \frac{\ln n}{K} \,, \tag{6.11}$$

the bound for the uniform allocation is at least

$$\Delta \exp \left(-\Delta^2 (4\alpha+1) \left(\frac{6+K}{4\Delta^2} \right) \frac{\ln n}{2K} \right) = \Delta n^{-(4\alpha+1)(6+K)/8K} \,,$$

which may be much worse than the upper bound (6.9) for the UCB strategy whenever K is large, as can be seen by comparing the exponents $-2(\alpha - 1)$ versus $-(4\alpha+1)(6+K)/8K$.

To illustrate this numerically (though this is probably not the most convincing choice of the parameters), consider the case when $\Delta = 0.4$, $K = 20$, and $\alpha = 4$. Then $n = 6020$ satisfies (6.10)

and (6.11). For these parameters, the upper bound (6.9) for the UCB strategy is 4.00×10^{-14}, which is much smaller than the one for the uniform allocation, which is larger than 1.45×10^{-11}.

The reason is that the uniform allocation strategy only samples $\lfloor n/K \rfloor$ each arm, whereas the UCB strategy focuses rapidly its exploration on the better arms.

CHAPTER 7

Pure Exploration in Multi-Armed Bandits II

We consider the problem of finding the best arm in a stochastic multi-armed bandit game. The regret of a forecaster is here defined by the gap between the mean reward of the optimal arm and the mean reward of the ultimately chosen arm. We propose a highly exploring UCB policy and a new algorithm based on successive rejects. We show that these algorithms are essentially optimal since their regret decreases exponentially at a rate which is, up to a logarithmic factor, the best possible. However, while the UCB policy needs the tuning of a parameter depending on the unobservable hardness of the task, the successive rejects policy benefits from being parameter-free, and also independent of the scaling of the rewards. As a by-product of our analysis, we show that identifying the best arm (when it is unique) requires a number of samples of order (up to a $\log(K)$ factor) $\sum_i 1/\Delta_i^2$, where the sum is on the suboptimal arms and Δ_i represents the difference between the mean reward of the best arm and the one of arm i. This generalizes the well-known fact that one needs of order of $1/\Delta^2$ samples to differentiate the means of two distributions with gap Δ.

Contents

This chapter is a joint work with Jean-Yves Audibert and Rémi Munos. It is based on the paper Audibert et al. [2010] published in the proceedings of the 23rd Annual Conference on Learning Theory.

1. Introduction

In the multi-armed bandit problem described in Chapter 2, at each stage, an agent (or forecaster) chooses one action (or arm), and receives a reward from it. In its stochastic version, the reward is drawn from a fixed probability distribution given the arm. The usual goal is to maximize the cumulative sum of rewards, see Robbins [1952], Auer et al. [2002] among many others. Since the forecaster does not know the distributions, he needs to explore (try) the different actions and yet, exploit (concentrate its draws on) the seemingly most rewarding arms. In this chapter, we

Parameters available to the forecaster: the number of rounds n and the number of arms K.

Parameters unknown to the forecaster: the reward distributions $\nu_1, \dots, \nu_K$ of the arms.

For each round $t = 1, 2, \dots, n$;

(1) the forecaster chooses $I_t \in \{1, \dots, K\}$,

(2) the environment draws the reward $X_{I_t, T_{I_t}(t)}$ from ν_{I_t} and independently of the past given I_t.

At the end of the n rounds, the forecaster outputs a recommendation $J_n \in \{1, \dots, K\}$.

Figure 1: The pure exploration problem for multi-armed bandits.

adopt a different viewpoint, which we already investigated in Chapter 6. We assume that after a given number of pulls, the forecaster is asked to output a recommended arm. He is then *only* evaluated by the average payoff of his recommended arm. This is the so-called pure exploration problem.

The distinguishing feature from the classical multi-armed bandit problem described above is that the exploration phase and the evaluation phase are separated. Thus, there is no explicit trade-off between the exploration and the exploitation while pulling the arms. The target of Hoeffding and Bernstein races, see Maron and Moore [1993], Mnih et al. [2008] among others, is more similar to ours. However, instead of trying to extract from a fixed number of rounds the best action, racing algorithms try to identify the best action at a given confidence level while consuming the minimal number of pulls. They optimize the budget for a given confidence level, instead of optimizing the quality of the recommendation for a given budget size.

In addition to the applications described in Chapter 6 for this framework, we propose another motivating with channel allocation for mobile phone communications. During a very short time before the communication starts, a cellphone can explore the set of channels to find the best one to operate. Each evaluation of a channel is noisy and there is a limited number of evaluations before the communication starts. The connection is then launched on the channel which is believed to be the best. Opportunistic communication systems rely on the same idea. Again the cumulative regret during the exploration phase is irrelevant since the user is only interested in the quality of its communication starting after the exploration phase.

2. Problem setup

A stochastic multi-armed bandit game is parameterized by the number of arms K, the number of rounds (or budget) n, and K probability distributions $\nu_1, \dots, \nu_K$ associated respectively with arm 1, …, arm K. These distributions are unknown to the forecaster. For $t = 1, \dots, n$, at round t, the forecaster chooses an arm I_t in the set of arms $\{1, \dots, K\}$, and observes a reward drawn from ν_{I_t} independently from the past (actions and observations). At the end of the n rounds, the forecaster selects an arm, denoted J_n, and is evaluated in terms of the difference between the mean reward of the optimal arm and the mean reward of J_n. Precisely, let $\mu_1, \dots, \mu_K$ be the respective means of $\nu_1, \dots, \nu_K$. Let $\mu^* = \max_{k \in \{1,\dots,K\}} \mu_k$. The simple regret of the forecaster is

$$r_n = \mu^* - \mu_{J_n}.$$

For sake of simplicity, *we will assume that the rewards are in* $[0,1]$ *and that there is a unique optimal arm*. Let i^* denote this arm (so, $\mu_{i^*} = \mu^*$). For $i \neq i^*$, we introduce the following suboptimality measure of arm i:

$$\Delta_i = \mu^* - \mu_i.$$

For reasons that will be obvious later, we also define Δ_{i^*} as the minimal gap

$$\Delta_{i^*} = \min_{i \neq i^*} \Delta_i.$$

We introduce the notation $(i) \in \{1, \dots, K\}$ to denote the i–th best arm (with ties break arbitrarily), hence

$$\Delta_{i^*} = \Delta_{(1)} = \Delta_{(2)} \leqslant \Delta_{(3)} \leqslant \ldots \leqslant \Delta_{(K)}.$$

Let e_n denote the probability of error, that is the probability that the recommendation is not the optimal one:

$$e_n = \mathbb{P}(J_n \neq i^*).$$

We have $\mathbb{E}r_n = \sum_{i \neq i^*} \mathbb{P}(J_n = i)\Delta_i$, and consequently

$$\Delta_{i^*} e_n \leqslant \mathbb{E}r_n \leqslant e_n.$$

As a consequence of this equation, up to a second order term, e_n and $\mathbb{E}r_n$ behave similarly, and it does not harm to focus on the probability e_n.

For each arm i and all time rounds $t \geqslant 1$, we denote by $T_i(t)$ the number of times arm i was pulled from rounds 1 to t, and by $X_{i,1}, X_{i,2}, \dots, X_{i,T_i(t)}$ the sequence of associated rewards. Introduce $\widehat{\mu}_{i,s} = \frac{1}{s}\sum_{t=1}^{s} X_{i,t}$ the empirical mean of arm i after s pulls. In the following, the symbol c will denote a positive numerical constant which may differ from line to line.

The goal of this work is to propose allocation strategies with small simple regret, and possibly as small as the best allocation strategy which would know beforehand the distributions $\nu_1, \dots, \nu_K$ up to a permutation. Before going further, note that the goal is unachievable for all distributions $\nu_1, \dots, \nu_K$: a policy cannot perform as well as the "oracle" allocation strategy in every particular cases. For instance, when the supports of $\nu_1, \dots, \nu_K$ are disjoints, the oracle forecaster almost surely identifies an arm by a single draw of it. As a consequence, it has almost surely zero simple regret for any $n \geqslant K$. The generic policy which does not have any knowledge on the K distributions cannot reproduce this performance for any K-tuple of disjointly supported distributions. In this work, the above goal of deciding as well as an oracle will be reached for the set of Bernoulli distributions with parameters in $(0,1)$, but the algorithms are defined for any distributions supported in $[0,1]$.

We would like to mention that the case $K = 2$ is unique and simple since, as we will indirectly see, it is optimally solved by the uniform allocation strategy consisting in drawing each arm $n/2$ times (up to rounding problem), and at the end recommending the arm with the highest empirical mean. Therefore, our main contributions concern more the problem of the budget allocation when $K \geqslant 3$. The hardness of the task will be characterized by the following quantities

$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2} \qquad \text{and} \qquad H_2 = \max_{i \in \{1,\dots,K\}} i\Delta_{(i)}^{-2}.$$

These quantities are equivalent up to a logarithmic factor since we have (see Section 8.1)

$$H_2 \leqslant H_1 \leqslant \log(2K) H_2. \tag{7.1}$$

Intuitively, we will show that these quantities are indeed characteristic of the hardness of the problem, in the sense that they give the order of magnitude of the number of samples required to find the best arm with a reasonable probability. This statement will be made precise in the rest of the

chapter, in particular through Theorem 7.2 and Theorem 7.4.

Outline. In Section 3, we propose a highly exploring policy based on upper confidence bounds, called UCB-E (Upper Confidence Bound Exploration), in the spirit of UCB1 Auer et al. [2002]. We prove that this algorithm, provided that it is appropriately tuned, has an upper bound on the probability of error e_n of order $\exp\big(-c\frac{n}{H_1}\big)$. The core problem of this policy is the tuning of the parameter. The optimal value of the parameter depends on H_1, which has no reason to be known beforehand by the forecaster, and which, to our knowledge, cannot be estimated from past observations with sufficiently high confidence in order that the resulting algorithm still satisfies a similar bound on e_n.

To get round this limitation, in Section 4, we propose a simple new policy called SR (Successive Rejects) that progressively rejects the arms which seem to be suboptimal. This algorithm is parameter-free and its probability of error e_n is at most of order $\exp\big(-\frac{n}{\log(2K)H_2}\big)$. Since $H_2 \leqslant H_1 \leqslant \log(2K)H_2$, up to at most a logarithmic term in K, the algorithm performs as well as UCB-E while not requiring the knowledge of H_1.

In Section 5, we prove that H_1 and H_2 truly represent the hardness of the problem (up to a logarithmic factor). Precisely, we consider a forecaster which knows the reward distributions of the arms *up to a permutation*. When these distributions are of Bernoulli type with parameter in $[p, 1-p]$ for some $p > 0$, there exists a permutation of the distributions for which the probability of error of the (oracle) forecaster is lower bounded by $\exp\big(-\frac{cn}{p(1-p)H_2}\big)$.

Section 6 provides some experiments testing the efficiency of the proposed policies and enlightening our theoretical results. We also discuss a modification of UCB-E where we perform a non-trivial online estimation of H_1. We conclude in Section 7. Finally section 8 gathers the proofs.

Example. To put in perspective the results we just mentioned, let us consider a specific example with Bernoulli distributions. Let $\nu_1 = \mathrm{Ber}\big(\frac{1}{2}\big)$, and $\nu_i = \mathrm{Ber}\big(\frac{1}{2} - \frac{1}{K^i}\big)$ for $i \in \{2, \ldots, K\}$. Here, one can easily check that $H_2 = 2K^{2K}$. Thus, in this case, the probability of missing the best arm of SR is at most of order $\exp\big(-\frac{n}{2\log(2K)K^{2K}}\big)$. Moreover, in Section 5, we prove that there does not exist any forecaster (even with the knowledge of the distributions up to a permutation) with a probability of missing the best arm smaller than $\exp\big(-\frac{11n}{K^{2K}}\big)$ for infinitely many n. Thus, our analysis finds that, for this particular reward distributions, the number of samples required to find the best arm is at least (of order of) K^{2K}, and SR actually finds it with (of order of) $\log(K)K^{2K}$ samples.

3. Highly exploring policy based on upper confidence bounds

In this section, we propose and study the algorithm UCB-E described in Figure 2. When a is taken of order $\log n$, the algorithm essentially corresponds to the UCB policy described in Section 2.2, and its cumulative regret is of order $\log n$. In Chapter 6 we showed that algorithms having at most logarithmic cumulative regret, has at least a simple regret of order $n^{-\gamma}$ for some $\gamma > 0$. So taking a of order $\log n$ is inappropriate to reach exponentially small probability of error. For the simple regret, one has to explore much more and typically use a parameter which is essentially linear in n. Precisely, we have the following result, which proof can be found in Section 8.2.

THEOREM 7.1. *If UCB-E is run with parameter* $0 < a \leqslant \frac{25}{36}\frac{n-K}{H_1}$*, then it satisfies*

$$e_n \leqslant 2nK \exp\left(-\frac{2a}{25}\right).$$

Parameter: exploration parameter $a > 0$.

For $i \in \{1, \dots, K\}$, let $B_{i,s} = \widehat{\mu}_{i,s} + \sqrt{\frac{a}{s}}$ for $s \geqslant 1$ and $B_{i,0} = +\infty$.

For each round $t = 1, 2, \dots, n$:

Draw $I_t \in \operatorname{argmax}_{i \in \{1,\dots,K\}} B_{i,T_i(t-1)}$.

Let $J_n \in \operatorname{argmax}_{i \in \{1,\dots,K\}} \widehat{\mu}_{i,T_i(n)}$.

Figure 2: UCB-E (Upper Confidence Bound Exploration) algorithm.

In particular for $a = \frac{25}{36}\frac{n-K}{H_1}$*, we have* $e_n \leqslant 2nK \exp\big(-\frac{n-K}{18H_1}\big)$.

The theorem shows that the probability of error of UCB-E is at most of order $\exp(-ca)$ for $a \geqslant \log n$. In fact, Theorem 7.5 in Appendix 8.3 shows a corresponding lower bound. In view of this, as long as $a \leqslant \frac{25}{36}\frac{n-K}{H_1}$, we can essentially say: the more we explore (i.e., the larger a is), the smaller the simple regret is. Besides, the smallest upper bound on the probability of error is obtained for a of order n/H_1, and is therefore exponentially decreasing with n. The constant H_1 depends not only on how close the mean rewards of the two best arms are, but also on the number of arms and how close their mean reward is to the optimal mean reward. This constant should be seen as the order of the minimal number n for which the recommended arm is the optimal one with high probability. In Section 5, we will show that H_1 is indeed a good measure of the hardness of the task by showing that no forecaster satisfies $e_n \leqslant \exp\big(-\frac{cn}{H_2}\big)$ for any distributions $\nu_1, \dots, \nu_K$, where we recall that H_2 satisfies $H_2 \leqslant H_1 \leqslant \log(2K)H_2$.

One interesting message to take from the proof of Theorem 7.1 is that, with probability at least $1 - 2nK \exp\big(-\frac{2a}{25}\big)$, the number of draws of any suboptimal arm i is of order $a\Delta_i^{-2}$. This means that the optimal arm will be played at least $n - caH_1$, showing that for too small a, UCB-E "exploits" too much in view of the simple regret. Theorem 7.1 does not specify how the algorithm performs when a is larger than $\frac{25}{36}\frac{n-K}{H_1}$. Nevertheless, similar arguments than the ones in the proof show that for large a, with high probability, only low rewarding arms are played of order $a\Delta_i^{-2}$ times, whereas the best ones are all drawn the same number of times up to a constant factor. The number of these similarly drawn arms grows with a. In the limit, when a goes to infinity, UCB-E is exactly the uniform allocation strategy studied in Chapter 6. In general[1], the uniform allocation has a probability of error which can be lower and upper bounded by a quantity of the form $\exp\big(-c\frac{n\Delta_{i^*}^2}{K}\big)$. It consequently performs much worse than UCB-E for $a = \frac{25}{36}\frac{n-K}{H_1}$, since $H_1 \leqslant K\Delta_{i^*}^{-2}$, and potentially $H_1 \ll K\Delta_{i^*}^{-2}$ for very large number of arms with heterogeneous mean rewards.

One straightforward idea to cope with the absence of an oracle telling us the value of H_1 would be to estimate online the parameter H_1 and use this estimation in the algorithm. Unfortunately, we were not able to prove, and do not believe that, this modified algorithm generally attains the expected rate of convergence. Indeed, overestimating H_1 leads to low exploring, and in the event when the optimal arm has given abnormally low rewards, the arm stops being drawn by the policy, its estimated mean reward is thus not corrected, and the arm is finally not recommended by the policy. On the contrary, underestimating H_1 leads to draw too much the suboptimal arms,

[1]We say "in general" to rule out some trivial cases (like when the reward distributions are all Dirac distributions) in which the probability of error e_n would be much smaller.

Let $A_1 = \{1, \ldots, K\}$, $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $n_0 = 0$ and for $k \in \{1, \ldots, K-1\}$,

$$n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil .$$

For each phase $k = 1, 2, \ldots, K-1$:

(1) For each $i \in A_k$, select arm i during $n_k - n_{k-1}$ rounds.

(2) Let $A_{k+1} = A_k \setminus \arg\min_{i \in A_k} \widehat{\mu}_{i,n_k}$ (we only remove one element from A_k, if there is a tie, select randomly the arm to dismiss among the worst arms).

Let J_n be the unique element of A_K.

Figure 3: SR (Successive Rejects) algorithm.

precluding a sufficiently accurate estimation of the mean rewards of the best arms. For this last case, things are in fact much more subtle than what can be retranscribed in these few lines, and we notice that keeping track of a lower bound on H_1 would lead to the correct rate only under appropriate assumptions on the decrease of the sequence $\Delta_{(k)}$, $k \in \{1, \ldots, K\}$. In Section 6 we push this idea and propose a way to estimate online H_1, however we solely justify the corresponding algorithm by experiments. In the next section we propose an algorithm which does not suffer from these limitations.

4. Successive Rejects algorithm

In this section, we describe and analyze a new algorithm, SR (Successive Rejects), see Figure 3 for its precise description. Informally it proceeds as follows. First the algorithm divides the time (i.e., the n rounds) in $K-1$ phases. At the end of each phase, the algorithm dismisses the arm with the lowest empirical mean. During the next phase, it pulls equally often each arm which has not been dismissed yet. The recommended arm J_n is the last surviving arm. The length of the phases are carefully chosen to obtain an optimal (up to a logarithmic factor) convergence rate. More precisely, one arm is pulled $n_1 = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K} \right\rceil$ times, one $n_2 = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K-1} \right\rceil$ times, ..., and two arms are pulled $n_{K-1} = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{2} \right\rceil$ times. SR does not exceed the budget of n pulls, since, from the definition $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, we have

$$n_1 + \ldots + n_{K-1} + n_{K-1} \leqslant K + \frac{n-K}{\overline{\log}(K)} \left(\frac{1}{2} + \sum_{k=1}^{K-1} \frac{1}{K+1-k} \right) = n.$$

For $K = 2$, up to rounding effects, SR is just the uniform allocation strategy.

THEOREM 7.2. *The probability of error of SR satisfies*

$$e_n \leqslant \frac{K(K-1)}{2} \exp\left(-\frac{n-K}{\overline{\log}(K) H_2} \right).$$

PROOF. We can assume that the sequence of rewards for each arm is drawn before the beginning of the game. Thus the empirical reward for arm i after s pulls is well defined even if arm i has not been actually pulled s times. During phase k, at least one of the k worst arms is surviving. So, if the optimal arm i^* is dismissed at the end of phase k, it means that $\widehat{\mu}_{i^*,n_k} \leqslant \max_{i \in \{(K),(K-1),\ldots,(K+1-k)\}} \widehat{\mu}_{i,n_k}$. By a union bound and Hoeffding's inequality, the

probability of error $e_n = \mathbb{P}(A_K \neq \{i^*\})$ thus satisfies

$$\begin{aligned} e_n &\leqslant \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}(\widehat{\mu}_{i^*,n_k} \leqslant \widehat{\mu}_{(i),n_k}) \\ &\leqslant \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}(\widehat{\mu}_{i^*,n_k} - \mu^* + \mu_{(i)} - \widehat{\mu}_{(i),n_k} \geqslant \Delta_{(i)}) \\ &\leqslant \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \exp\left(-n_k \Delta_{(i)}^2\right) \leqslant \sum_{k=1}^{K-1} k \exp\left(-n_k \Delta_{(K+1-k)}^2\right). \end{aligned}$$

We conclude the proof by noting that by definition of n_k and H_2, we have

$$n_k \Delta_{(K+1-k)}^2 \geqslant \frac{n-K}{\overline{\log}(K)} \frac{1}{(K+1-k)\Delta_{(K+1-k)}^{-2}} \geqslant \frac{n-K}{\overline{\log}(K) H_2}. \tag{7.2}$$

□

The following theorem provides a deeper understanding of how SR works. It lower bounds the sampling times of the arms and shows that at the end of phase k, we have a high-confidence estimation of $\Delta_{(K+1-k)}$ up to numerical constant factor. This intuition will prove to be useful in Section 6, see in particular Figure 4.

THEOREM 7.3. *With probability at least* $1 - \frac{K^3}{2}\exp\big(-\frac{n-K}{4\overline{\log}(K)H_2}\big)$*, for any arm* j*, we have*

$$T_j(n) \geqslant \frac{n-K}{4\overline{\log}(K) H_2 \Delta_j^2}. \tag{7.3}$$

With probability at least $1 - K^3\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$*, for any* $k \in \{1,\dots,K-1\}$*, the dismissed arm* $\ell_k = A_{k+1} \setminus A_k$ *at the end of phase* k *satisfies*

$$\frac{1}{4}\Delta_{(K+1-k)} \leqslant \frac{1}{2}\Delta_{\ell_k} \leqslant \max_{m \in A_k} \widehat{\mu}_{m,n_k} - \widehat{\mu}_{\ell_k,n_k} \leqslant \frac{3}{2}\Delta_{\ell_k} \leqslant 3\Delta_{(K+1-k)}. \tag{7.4}$$

PROOF. We consider the event $\mathcal{E}$ on which for any $k \in \{1,\dots,K-1\}$, for any arm ℓ in the worst k arms, and any arm j such that $2\Delta_j \leqslant \Delta_\ell$, we have

$$\widehat{\mu}_{j,n_k} - \widehat{\mu}_{\ell,n_k} > 0.$$

This event holds with probability at least $1 - \frac{K^3}{2}\exp\big(-\frac{n-K}{4\overline{\log}(K)H_2}\big)$, since, from Hoeffding's inequality, a union bound and (7.2), we have

$$\begin{aligned} &\sum_{k=1}^{K-1} \sum_{\substack{\ell \in \{(K),(K-1),\dots,(K+1-k)\} \\ j:2\Delta_j \leqslant \Delta_\ell}} \mathbb{P}\Big(\widehat{\mu}_{j,n_k} - \widehat{\mu}_{\ell,n_k} \leqslant 0\Big) \\ &\leqslant \sum_{k=1}^{K-1} \sum_{\substack{\ell \in \{(K),(K-1),\dots,(K+1-k)\} \\ j:2\Delta_j \leqslant \Delta_\ell}} \exp\big(-n_k(\Delta_\ell - \Delta_j)^2\big) \\ &\leqslant \sum_{k=1}^{K-1} kK \exp\big(-n_k \frac{\Delta_{(K+1-k)}^2}{4}\big) \leqslant \frac{K^3}{2}\exp\left(-\frac{n-K}{4\overline{\log}(K)H_2}\right). \end{aligned}$$

During phase k, at least one of the k worst arms is surviving. On the event $\mathcal{E}$, this surviving arm has an empirical mean at the end of the phase which is smaller than the one of any arm j satisfying $2\Delta_j \leqslant \Delta_{(K+1-k)}$. So, at the end of phase k, any arm j satisfying $2\Delta_j \leqslant \Delta_{(K+1-k)}$ cannot

be dismissed. Now, for a given arm j, we consider two cases depending whether there exists $m \in \{1, \dots, K\}$ such that $\Delta_{(m-1)} \leqslant 2\Delta_j \leqslant \Delta_{(m)}$.

First case. If no such m exists, then we have $\Delta_j^2 T_j(n) \geqslant \frac{1}{4}\Delta_{(K)}^2 n_1 \geqslant \frac{n-K}{4\overline{\log}(K)H_2}$, so that (7.3) holds.

Second case. If such m exists, then, from the above argument, the arm j cannot be dismissed before the end of the phase $K+2-m$ (since there exists $K+1-m$ arms ℓ such that $\Delta_\ell \geqslant 2\Delta_j$). From (7.2), we get

$$\Delta_j^2 T_j(n) \geqslant \Delta_j^2 n_{K+2-m} \geqslant \frac{\Delta_j^2}{\Delta_{(m-1)}^2} \frac{n-K}{\overline{\log}(K)H_2} \geqslant \frac{n-K}{4\overline{\log}(K)H_2},$$

which ends the proof of (7.3). We have seen that at the end of phase k, any arm j satisfying $2\Delta_j \leqslant \Delta_{(K+1-k)}$ cannot be dismissed. Consequently, at the end of phase k, the dismissed arm $\ell_k = A_{k+1} \setminus A_k$ satisfies the left inequality of

$$\frac{1}{2}\Delta_{(K+1-k)} \leqslant \Delta_{\ell_k} \leqslant 2\Delta_{(K+1-k)}. \tag{7.5}$$

Let us now prove the right inequality by contradiction. Consider k such that $2\Delta_{(K+1-k)} < \Delta_{\ell_k}$. Arm ℓ_k thus belongs to the $k-1$ worst arms. Hence, in the first $k-1$ rejects, say at the end of phase k', an arm j with $\Delta_j \leqslant \Delta_{(K+1-k)}$ is dismissed. From the left inequality of (7.5), we get $\Delta_{(K+1-k')} \leqslant 2\Delta_j < \Delta_{\ell_k}$. On the event $\mathcal{E}$, we thus have $\widehat{\mu}_{j,n_{k'}} - \widehat{\mu}_{\ell_k,n_{k'}} > 0$ (since ℓ_k belongs to the k' worst arms by the previous inequality). This contradicts the fact that j is rejected at phase k'. So (7.5) holds.

Now let $\mathcal{E}'$ be the event on which for any arm j, and any $k \in \{1, \dots, K-1\}$ $\left|\widehat{\mu}_{j,n_k} - \mu_j\right| \leqslant \frac{\Delta_{(K+1-k)}}{8}$. Using again Hoeffding's inequality, a union bound and (7.2), this event holds with probability at least $1 - 2K(K-1)\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$. We now work on the event $\mathcal{E} \cap \mathcal{E}'$, which holds with probability at least $1 - K^3\exp\big(-\frac{n-K}{32\overline{\log}(K)H_2}\big)$. From (7.5), the dismissed arm ℓ_k at the end of phase k satisfies

$$\left|\widehat{\mu}_{\ell_k,n_k} - \mu_{\ell_k}\right| \leqslant \frac{\Delta_{(K+1-k)}}{8} \leqslant \frac{\Delta_{\ell_k}}{4}.$$

Besides, we also have

$$\Big|\max_{m \in A_k} \widehat{\mu}_{m,n_k} - \mu_{(1)}\Big| \leqslant \frac{\Delta_{(K+1-k)}}{8} \leqslant \frac{\Delta_{\ell_k}}{4}.$$

Consequently, at the end of phase k, we have

$$\frac{1}{4}\Delta_{(K+1-k)} \leqslant \frac{1}{2}\Delta_{\ell_k} \leqslant \max_{m \in A_k} \widehat{\mu}_{m,n_k} - \widehat{\mu}_{\ell_k,n_k} \leqslant \frac{3}{2}\Delta_{\ell_k} \leqslant 3\Delta_{(K+1-k)}.$$

□

5. Lower bound

In this section we provide a very general and somewhat surprising lower bound. We prove that, when the reward distributions are Bernoulli distributions with variances bounded away from 0, then for any forecaster, one can permute the distributions on the arms (before the game starts) so that the probability of missing the best arm will be at least of order $\exp\left(-\frac{cn}{H_2}\right)$. Note that, in this formulation, we allow the forecaster to *know* the reward distributions up to a permutation of the indexes! However, as the lower bound expresses it, even in this relatively easier case, the quantity H_2 is a good measure of the hardness of finding the best arm

THEOREM 7.4 (Lower Bound). *Let $\nu_1, \dots, \nu_K$ be Bernoulli distributions with parameters in $[p, 1-p]$, $p \in (0, 1/2)$. For any forecaster, there exists a permutation $\sigma : \{1, \dots, K\} \to \{1, \dots, K\}$ such that the probability error of the forecaster on the bandit problem defined by $\widetilde{\nu}_1 = \nu_{\sigma(1)}, \dots, \widetilde{\nu}_K = \nu_{\sigma(K)}$ satisfies*

$$e_n \geqslant \exp\left(-\frac{(5+o(1))n}{p(1-p)H_2}\right),$$

where the $o(1)$ term depends only on K, p and n and goes to 0 when n goes to infinity (see the end of the proof).

The proof of this result is quite technical. However, it is simple to explain why we can expect such a bound to hold. Assume (without loss of generality) that the arms are ordered, i.e., $\mu_1 > \mu_2 \geqslant \ldots \geqslant \mu_K$, and that all rewards $X_{i,t}$, $t \in \{1, \dots, n\}$, $i \in \{1, \dots, K\}$, are drawn before the game starts. Let $i \in \{2, \dots, K\}$. If $\widehat{X}_{1,n/i} < \widehat{X}_{i,n/i} \leqslant \widehat{X}_{j,n/i}$ for all $j \in \{2, \dots, i-1\}$, then it seems reasonable that a good forecaster should not pull arm 1 more than n/i times, and furthermore not select it as its recommendation. One can see that, the probability of the event we just described is of order of $\exp(-c(n/i)\Delta_i^2)$. Thus, with probability at least $\exp(-cn/\max_{2 \leqslant i \leqslant K} i\Delta_i^{-2})$, the forecaster makes an error, which is exactly the lower bound we propose. However, note that this argument does not yield a reasonable proof strategy, in particular we assumed a "good" forecaster with a "reasonable" behavior. For instance, it is obvious that the proof has to permute the arms, since a forecaster could, despite all, choose arm 1 as its recommendation, which imply a probability error of 0 as soon as the best arm is in position 1.

The main idea of our proposed proof goes as follows. A bandit problem is defined by a product distribution $\nu = \nu_1 \otimes \cdots \otimes \nu_K$. One can consider that at the beginning of the game, n K-tuples of rewards is sampled from this product distribution. This defines a table of nK rewards. A forecaster will explore a subpart of this table. We want to find a permutation σ of $\{1, \dots, K\}$ so that the indices of the best arm for ν and $\widetilde{\nu} = \nu_{\sigma(1)} \otimes \cdots \otimes \nu_{\sigma(K)}$ are different and such that the likelihood ratio of the explored part of the table of nK rewards under ν and $\widetilde{\nu}$ is at least of order $\exp(-cn/H_2)$ with probability with respect to $\nu^{\otimes n}$ lower bounded by a positive numerical constant. This would imply the claimed bound. Remark that, the "likelihood cost" of moving distribution ν_i to arm j depends on both the (Kullback-Leibler) distance between the distributions ν_i and ν_j, and the number of times arm j is pulled. Thus, we have to find the right trade-off between moving a distribution to a "close" distribution, and the fact that the target arm should not be pulled too much. To do this, we "slice" the set of indexes in a non-trivial (and non-intuitive) way. This "slicing" depends only on the reward distributions, and not on the considered forecaster. Then, to put it simply, we move the less drawn arm from one slice to the less drawn arm in the next slice. Note that the preceding sentence is not well defined, since by doing this we would get a random permutation (which of course does not make sense to derive a lower bound). However, at the cost of some technical difficulties, it is possible to circumvent this issue.

To achieve the program outlined above, as already hinted, we use the Kullback-Leibler divergence, which is defined for two probability distributions ρ, ρ' on $[0, 1]$ with ρ absolutely continuous with respect to ρ' as:

$$\mathrm{KL}(\rho, \rho') = \int_0^1 \log\left(\frac{d\rho}{d\rho'}(x)\right) d\rho(x) = \mathbb{E}_{X \sim \rho} \log\left(\frac{d\rho}{d\rho'}(X)\right).$$

Another quantity of particular interest for our analysis is

$$\widehat{\mathrm{KL}}_{i,t}(\rho,\rho') = \sum_{s=1}^{t} \log\left(\frac{d\rho}{d\rho'}(X_{i,s})\right).$$

In particular note that, if arm i has distribution ρ, then this quantity represents the (non re-normalized) empirical estimation of $\mathrm{KL}(\rho,\rho')$ after t pulls of arm i. Let $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ the probability and expectation signs when we integrate with respect to the distribution $\nu^{\otimes n}$. Another important property is that for any two product distributions ν,ν', which differ only on index i, and for any event A, one has:

$$\mathbb{P}_\nu(A) = \mathbb{E}_{\nu'}\mathbb{1}_A \exp\left(-\widehat{\mathrm{KL}}_{i,T_i(n)}(\nu_i',\nu_i)\right), \tag{7.6}$$

since we have $\prod_{s=1}^{T_{i,n}} \frac{d\nu_i}{d\nu_i'}(X_{i,s}) = \exp\big(-\widehat{\mathrm{KL}}_{i,T_i(n)}(\nu_i',\nu_i)\big)$.

PROOF. **First step: Notations.** Without loss of generality we can assume that ν is ordered in the sense that $\mu_1 > \mu_2 \geqslant \ldots \geqslant \mu_K$. Moreover let $L \in \{2,\ldots,K\}$ such that $H_2 = L/\Delta_L^2$, that is for all $i \in \{1,\ldots,K\}$,

$$i/\Delta_i^2 \leqslant L/\Delta_L^2. \tag{7.7}$$

We define now recursively the following sets. Let $k_1 = 1$,

$$\Sigma_1 = \left\{i : \mu_L \leqslant \mu_i \leqslant \mu_L + \frac{\Delta_L}{L^{1/2^{k_1}}}\right\},$$

and for $j > 1$,

$$\Sigma_j = \left\{i : \mu_L + \frac{\Delta_L}{L^{1/2^{k_{j-1}}}} < \mu_i \leqslant \mu_L + \frac{\Delta_L}{L^{1/2^{k_j}}}\right\},$$

where k_j is the smallest integer (if it exists, otherwise set $k_j = +\infty$) such that $|\Sigma_j| > 2|\Sigma_{j-1}|$. Let $\ell = \max\{j : k_j < +\infty\}$. We define now the random variables $Z_1,\ldots,Z_\ell$ corresponding to the indices of the less sampled arms of the respective slices $\Sigma_1,\ldots,\Sigma_\ell$: for $j \in \{1,\ldots,\ell\}$,

$$Z_j \in \operatorname*{argmin}_{i\in\Sigma_j} T_i(n).$$

Finally let $Z_{\ell+1} \in \operatorname{argmin}_{i\in\{1,\ldots,L\}\setminus\{J_n\}} T_i(n)$.

Second step: Controlling $T_{Z_j}(n)$, $j \in \{1,\ldots,\ell+1\}$. We first prove that for any $j \in \{1,\ldots,\ell\}$,

$$3|\Sigma_j| \geqslant L^{1-\frac{1}{2^{k_{j+1}-1}}}. \tag{7.8}$$

To do so let us note that, by definition of k_{j+1}, we have

$$\begin{aligned} 2|\Sigma_j| &\geqslant \left|\left\{i : \mu_L + \Delta_L/L^{1/2^{k_j}} < \mu_i \leqslant \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right| \\ &\geqslant \left|\left\{i : \mu_i \leqslant \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right| - (|\Sigma_1| + \ldots + |\Sigma_{j-1}|). \end{aligned}$$

Now remark that, by definition again, we have $|\Sigma_1| + \ldots + |\Sigma_{j-1}| \leqslant (2^{-(j-1)} + \ldots + 2^{-1})|\Sigma_j| \leqslant |\Sigma_j|$. Thus we obtain $3|\Sigma_j| \geqslant \left|\left\{i : \mu_i \leqslant \mu_L + \Delta_L/L^{1/2^{k_{j+1}-1}}\right\}\right|$. We finish the proof of (7.8) with the following calculation, which makes use of (7.7). For any $v \geqslant 1$,

$$\begin{aligned} |\{i : \mu_i \leqslant \mu_L + \Delta_L/v\}| &= |\{i : \Delta_i \geqslant \Delta_L(1-1/v)\}| \\ &\geqslant \left|\left\{i : \sqrt{\frac{i}{L}}\Delta_L \geqslant \Delta_L(1-1/v)\right\}\right| \end{aligned}$$

$$= \big|\big\{i : i \geqslant L(1-1/v)^2\big\}\big| \geqslant L\bigg(1-(1-1/v)^2\bigg) \geqslant L/v.$$

Now (7.8) directly entails (since a minimum is smaller than an average), for $j \in \{1,\dots,\ell\}$,

$$T_{Z_j}(n) \leqslant 3L^{\frac{1}{2^{k_{j+1}-1}}-1} \sum_{i\in\Sigma_j} T_i(n). \tag{7.9}$$

Besides, since $Z_{\ell+1}$ is the less drawn arm among $L-1$ arms, we trivially have

$$T_{Z_{\ell+1}}(n) \leqslant \frac{n}{L-1}. \tag{7.10}$$

Third step: A change of measure. Let $\nu' = \nu_L \otimes \nu_2 \otimes \cdots \otimes \nu_K$ be a modified product distribution where we replaced the best distribution by ν_L. Now let us consider the event

$$C_n = \Big\{\forall t \in\{1,\dots,n\},\ i \in \{2,\dots,L\},\ j \in \{1,\dots,L\},$$
$$\widehat{\mathrm{KL}}_{i,t}(\nu_i,\nu_j) \leqslant t\ \mathrm{KL}(\nu_i,\nu_j) + o_n \qquad \text{and} \qquad \widehat{\mathrm{KL}}_{1,t}(\nu_L,\nu_j) \leqslant t\ \mathrm{KL}(\nu_L,\nu_j) + o_n\Big\},$$

where $o_n = 2\log(p^{-1})\sqrt{n\log(2L)}$. From Hoeffding's maximal inequality, we have $\mathbb{P}_{\nu'}(C_n) \geqslant 1/2$ (see Appendix 8.4). We thus have $\sum_{1\leqslant z_1,\dots,z_{\ell+1}\leqslant L} \mathbb{P}_{\nu'}\big(C_n \cap \{Z_1 = z_1,\dots,Z_{\ell+1} = z_{\ell+1}\}\big) \geqslant 1/2$. Moreover note that $Z_1,\dots,Z_\ell$ are all distinct. Thus there exists $\ell+1$ constants $z_1,\dots,z_{\ell+1}$ such that, for $A_n = C_n \cap \{Z_1 = z_1,\dots,Z_{\ell+1} = z_{\ell+1}\}$, we have

$$\mathbb{P}_{\nu'}(A_n) \geqslant \frac{1}{2L\times L!}. \tag{7.11}$$

Since, by definition $Z_{\ell+1} \neq J_n$, we have

$$A_n \subset \{J_n \neq z_{\ell+1}\}. \tag{7.12}$$

In the following we treat differently the cases $z_{\ell+1} = 1$ and $z_{\ell+1} \neq 1$. Let us assume in a first time that $z_{\ell+1} = 1$. Then, an application of (7.6) and (7.12) directly gives, by definition of A_n,

$$\begin{aligned}
e_n(\nu) = \mathbb{P}_\nu(J_n \neq 1) &= \mathbb{E}_{\nu'}\mathbb{1}_{J_n\neq 1}\exp\bigg(-\widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L,\nu_1)\bigg)\\
&\geqslant \mathbb{E}_{\nu'}\mathbb{1}_{A_n}\exp\bigg(-\widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L,\nu_1)\bigg)\\
&\geqslant \mathbb{E}_{\nu'}\mathbb{1}_{A_n}\exp\bigg(-o_n - T_{Z_{\ell+1}}(n)\mathrm{KL}(\nu_L,\nu_1)\bigg)\\
&\geqslant \frac{1}{2L\times L!}\exp\bigg(-o_n - \frac{n}{L-1}\mathrm{KL}(\nu_L,\nu_1)\bigg),
\end{aligned}$$

where we used (7.10) and (7.11) for the last equation. Now, note that Lemma 10.3 implies

$$\mathrm{KL}(\nu_L,\nu_1) \leqslant \frac{\Delta_L^2}{p(1-p)},$$

which concludes the proof in the case $z_{\ell+1} = 1$.

Assume now that $z_{\ell+1} \neq 1$. In this case we prove that the lower bound holds for a well defined permuted product distribution $\widetilde{\nu}$ of ν. We define it as follows. Let m be the smallest $j \in \{1,\dots,\ell+1\}$ such that $z_m = z_{\ell+1}$. Now we set $\widetilde{\nu}$ as follows: $\widetilde{\nu}_{z_m} = \nu_1$, $\widetilde{\nu}_{z_{m-1}} = \nu_{z_m}$, …, $\widetilde{\nu}_{z_1} = \nu_{z_2}$, $\widetilde{\nu}_1 = \nu_{z_1}$, and $\widetilde{\nu}_j = \nu_j$ for other values of j in $\{1,\dots,K\}$. Remark that $\widetilde{\nu}$ is indeed the result of a permutation of the distributions of ν. Again, an application of (7.6) and (7.12) gives, by definition of A_n,

$$
\begin{aligned}
e_n(\widetilde{\nu}) &= \mathbb{P}_{\widetilde{\nu}}(J_n \neq z_m) \\
&= \mathbb{E}_{\nu'} \mathbb{1}_{J_n \neq z_m} \exp\bigg(- \widehat{\mathrm{KL}}_{1,T_1(n)}(\nu_L, \nu_{z_1}) - \sum_{j=1}^{m-1} \widehat{\mathrm{KL}}_{z_j,T_{z_j}(n)}(\nu_{z_j}, \nu_{z_{j+1}}) - \widehat{\mathrm{KL}}_{z_m,T_{z_m}(n)}(\nu_{z_m}, \nu_{z_1}) \bigg) \\
&\geqslant \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\bigg(- (m+1) o_n - T_1(n) \mathrm{KL}(\nu_L, \nu_{Z_1}) - \sum_{j=1}^{m-1} T_{Z_j}(n) \mathrm{KL}(\nu_{Z_j}, \nu_{Z_{j+1}}) \\
&\qquad\qquad - T_{Z_m}(n) \mathrm{KL}(\nu_{Z_m}, \nu_{Z_1}) \bigg). \qquad (7.13)
\end{aligned}
$$

From Lemma 10.3, the definition of Σ_j, and since the parameters of the Bernoulli distributions are in $[p, 1-p]$, we have $\mathrm{KL}(\nu_L, \nu_{Z_1}) \leqslant \frac{1}{p(1-p)} \frac{\Delta_L^2}{L}$, $\mathrm{KL}(\nu_{Z_m}, \nu_{Z_1}) \leqslant \frac{\Delta_L^2}{p(1-p)}$, and for any $j \in \{1, \dots, m-1\}$,

$$
\mathrm{KL}(\nu_{Z_j}, \nu_{Z_{j+1}}) \leqslant \frac{1}{p(1-p)} \left(\frac{\Delta_L}{L^{1/2^{k_{j+1}}}} \right)^2 .
$$

Reporting these inequalities, as well as (7.9), (7.10) and (7.11) in (7.13), we obtain:

$$
\begin{aligned}
e_n(\widetilde{\nu}) &\geqslant \mathbb{E}_{\nu'} \mathbb{1}_{A_n} \exp\bigg(- (m+1) o_n - 3 \frac{\Delta_L^2}{p(1-p)L} \bigg(T_1(n) + \sum_{j=1}^{m-1} \sum_{i \in \Sigma_j} T_i(n) + \frac{nL}{3(L-1)} \bigg) \bigg) \\
&\geqslant \frac{1}{2L \times L!} \exp\bigg(- L\, o_n - 3n \frac{\Delta_L^2}{p(1-p)L} \bigg(1 + \frac{L}{3(L-1)} \bigg) \bigg)
\end{aligned}
$$

Since $L \leqslant K$ and $2K \times K! \leqslant \exp\big(2K \log(K)\big)$ and from the definitions of o_n and L, we obtain

$$
e_n(\widetilde{\nu}) \geqslant \exp\left(-2K \log(K) - 2K \log(p^{-1}) \sqrt{n \log(2K)} - 5 \frac{n}{p(1-p) H_2} \right),
$$

which concludes the proof. □

6. Experiments

We propose a few simple experiments to illustrate our theoretical analysis. As a baseline comparison we use the Hoeffding Race algorithm, see Maron and Moore [1993], and the uniform strategy, which pulls equally often each arm and recommend the arm with the highest empirical mean, see Section 4.1 for its theoretical analysis. We consider only Bernoulli distributions, and the optimal arm always has parameter $1/2$. Each experiment corresponds to a different situation for the gaps, they are either clustered in few groups, or distributed according to an arithmetic or geometric progression. In each experiment we choose the number of samples (almost) equal to H_1 (except for the last experiment where we run it twice, the second time with $2H_1$ samples). If our understanding of the meaning of H_1 is sound, in each experiment the strategies SR and UCB-E should be able to find the best arm with a reasonable probability (which should be roughly of the same order in each experiment). We report our results in Figure 5. The parameters for the experiments are as follows:

– Experiment 1: One group of bad arms, $K = 20$, $\mu_{2:20} = 0.4$ (meaning for any $j \in \{2, \dots, 20\}$, $\mu_j = 0.4$)

– Experiment 2: Two groups of bad arms, $K = 20$, $\mu_{2:6} = 0.42$, $\mu_{7:20} = 0.38$.

– Experiment 3: Geometric progression, $K = 4$, $\mu_i = 0.5 - (0.37)^i$, $i \in \{2, 3, 4\}$.

Parameter: exploration rate $c > 0$.

Definitions: For $k \in \{1, \ldots, K-1\}$, let $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil$, $t_0 = 0$, $t_1 = Kn_1$, and for $k > 1$, $t_k = n_1 + \ldots n_{k-1} + (K-k+1)n_k$.
For $i \in \{1, \ldots, K\}$ and $a > 0$, let $B_{i,s}(a) = \widehat{\mu}_{i,s} + \sqrt{\frac{a}{s}}$ for $s \geqslant 1$ and $B_{i,0} = +\infty$.

Algorithm: For each phase $k = 0, 1, \ldots, K-1$:
Let $\widehat{H}_{1,k} = K$ if $k = 0$, and otherwise

$$\widehat{H}_{1,k} = \max_{K-k+1 \leqslant i \leqslant K} i \widehat{\Delta}^{-2}_{<i>},$$

where $\widehat{\Delta}_i = \left(\max_{1 \leqslant j \leqslant K} \widehat{X}_{j,T_j(t_k)} \right) - \widehat{X}_{i,T_i(t_k)}$ and $< i >$ is an ordering such that $\widehat{\Delta}_{<1>} \leqslant \ldots \leqslant \widehat{\Delta}_{<K>}$.
For $t = t_k + 1, \ldots, t_{k+1}$:
Draw $I_t \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} B_{i,T_i(t-1)}(cn/\widehat{H}_{1,k})$.

Recommendation: Let $J_n \in \mathrm{argmax}_{i \in \{1,\ldots,K\}} \widehat{\mu}_{i,T_i(n)}$.

Figure 4: Adaptive UCB-E algorithm. Its intuitive justification goes as follows: The time points t_k correspond to the moments where the Successive Rejects algorithm would dismiss an arm. Intuitively, in light of Theorem 7.3, one can say that at time t_k a good algorithm should have reasonable approximation of the gaps between the best arm and the k worst arms, that is the quantities $\Delta_{(K-k+1)}, \ldots, \Delta_{(K)}$. Now with these quantities, one can build a lower estimate of H_2 and thus also of H_1. We use this estimate between the time points t_k and t_{k+1} to tune the parameter a of UCB-E.

– Experiment 4: 6 arms divided in three groups, $K = 6$, $\mu_2 = 0.42$, $\mu_{3:4} = 0.4$, $\mu_{5:6} = 0.35$.
– Experiment 5: Arithmetic progression, $K = 15$, $\mu_i = 0.5 - 0.025i$, $i \in \{2, \ldots, 15\}$.
– Experiment 6: Two good arms and a large group of bad arms, $K = 20$, $\mu_2 = 0.48$, $\mu_{3:20} = 0.37$.
– Experiment 7: Three groups of bad arms, $K = 30$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$.

The different graphics should be read as follows: Each bar represents a different algorithm and the bar's height represents the probability of error of this algorithm. The correspondence between algorithms and bars is the following:
– Bar 1: Uniform sampling strategy.
– Bar 2-4: Hoeffding Race algorithm with parameters $\delta = 0.01, 0.1, 0.3$.
– Bar 5: Successive Rejects strategy.
– Bar 6-9: UCB-E with parameter $a = cn/H_1$ where respectively $c = 1, 2, 4, 8$.
– Bar 10-14: Adaptive UCB-E (see Figure 4) with parameters $c = 1/4, 1/2, 1, 2, 4$.

7. Conclusion

This work has investigated strategies for finding the best arm in a multi-armed bandit problem. It has proposed a simple parameter-free algorithm, SR, that attains optimal guarantees up to a logarithmic term (Theorem 7.2 and Theorem 7.4). A precise understanding of both SR (Theorem 7.3) and a UCB policy (Theorem 7.1) lead us to define a new algorithm, Adaptive UCB-E. It comes without guarantee of optimal rates (see end of Section 3), but performs better than SR in practice (for $c = 1$, Adaptive UCB-E outperformed SR on all the experiments we did, even those done to

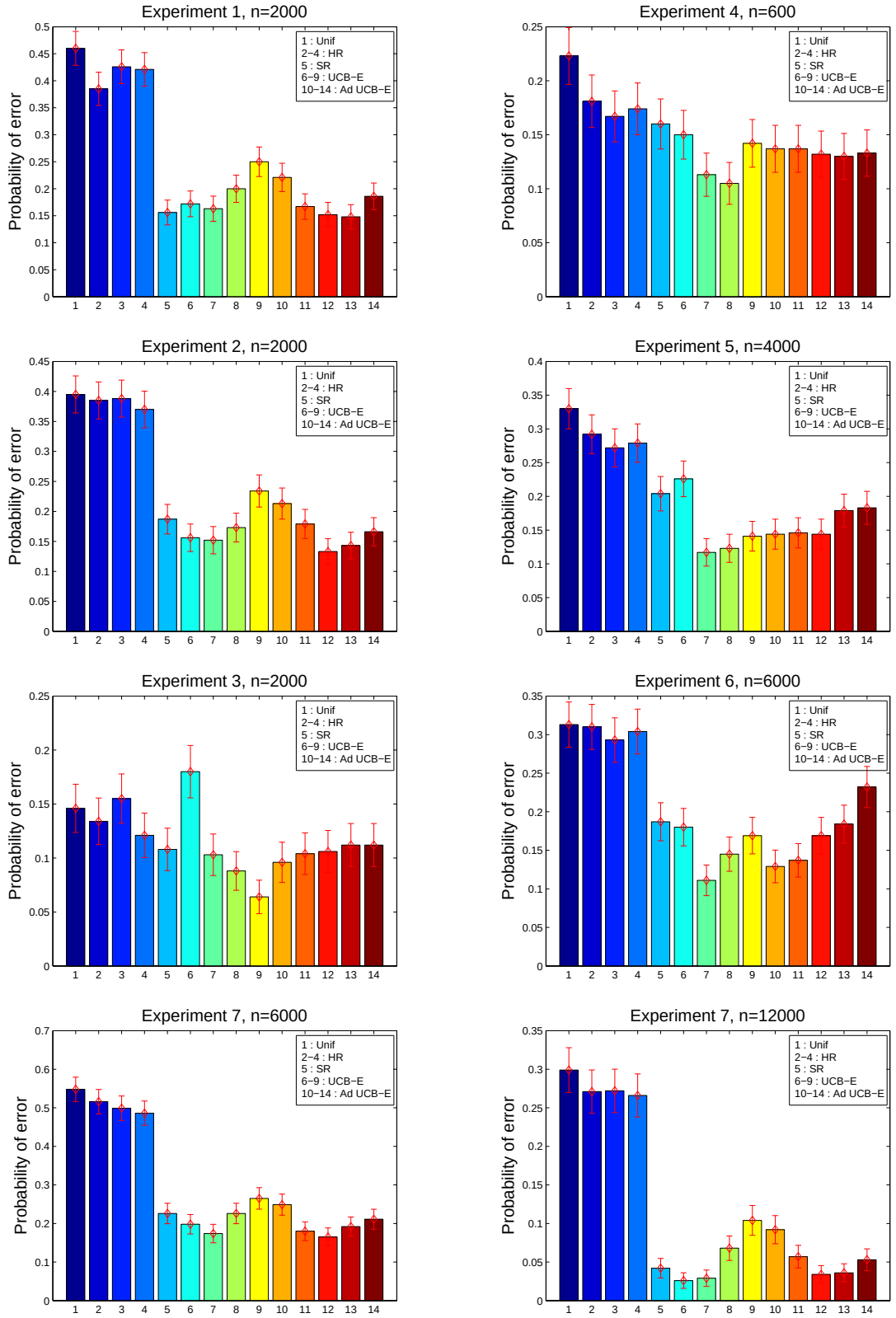

Figure 5: These results support our theoretical findings in the following sense: Despite the fact that the experiments are very different, one can see that since we use a number of samples (almost) equal to the hardness H_1, in all of them we get a probability of error of the same order, and moreover this probability is small enough to say that we identified the best arm. Note that the Successive Rejects algorithm represents in all cases a substantial improvement over both the naive uniform strategy and Hoeffding Race. These results also justify experimentally the algorithm Adaptive UCB-E, indeed one can see that with the constant $c = 1$ we obtain better results than SR in all experiments, even in experiment 6 which was designed to be a difficult instance of Adaptive UCB-E.

make it fail). One possible explanation is that SR is too static: it does not implement more data driven arguments such as: in a phase, a surviving arm performing much worse than the other ones is still drawn until the end of the phase even if it is clear that it is the next dismissed arm.

Extensions of this work may concentrate on the following problems. (i) What is a good measure of hardness when one takes into account the (empirical) variances? Do we have a good scaling with respect to the variance with the current algorithms or do we need to modify them ? (ii) Is it possible to derive a natural anytime version of Successive Rejects (without using a doubling trick)? (iii) Is it possible to close the logarithmic gap between the lower and upper bounds? (iv) How should we modify the algorithm and the analysis if one is interested in recommending the top m actions instead of a single one?

8. Proofs

8.1. Proof of Inequalities (7.1)**.** Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$. Remark that $\log(K+1) - 1/2 \leqslant \overline{\log}(K) \leqslant \log(K) + 1/2 \leqslant \log(2K)$. Precisely, we will prove

$$H_2 \leqslant H_1 \leqslant \overline{\log}(K)\ H_2,$$

which is tight to the extent that the right inequality is an equality when for some $0 < c \leqslant 1/\sqrt{K}$, we have $\Delta_{(i)} = \sqrt{i}c$ for any $i \neq i^*$, and the left inequality is an equality if all Δ_i's are equal.
Proof: The left inequality follows from: for any $i \in \{1, \ldots, K\}$,

$$H_1 = \sum_{k=1}^{K} \Delta_{(k)}^{-2} \geqslant \sum_{k=1}^{i} \Delta_{(i)}^{-2} \geqslant i\Delta_{(i)}^{-2}.$$

The right inequality directly comes from

$$\sum_{i=1}^{K} \Delta_{(i)}^{-2} = \Delta_{(2)}^{-2} + \sum_{i=2}^{K} \frac{1}{i} i\Delta_{(i)}^{-2} \leqslant \overline{\log}(K) \max_{i \in \{1,\ldots,K\}} i\Delta_{(i)}^{-2}.$$

8.2. Proof of Theorem 7.1. First step. Let us consider the event

$$\xi = \left\{ \forall i \in \{1, \ldots, K\}, s \in \{1, \ldots, n\}, |\widehat{\mu}_{i,s} - \mu_i| < \frac{1}{5}\sqrt{\frac{a}{s}} \right\}.$$

From Hoeffding's inequality and a union bound, we have $\mathbb{P}(\xi) \geqslant 1 - 2nK \exp\left(-\frac{2a}{25}\right)$. In the following, we prove that on the event ξ we have $J_n = i^*$, which concludes the proof. Since J_n is the empirical best arm, and given that we are on ξ, it is enough to prove that

$$\frac{1}{5}\sqrt{\frac{a}{T_i(n)}} \leqslant \frac{\Delta_i}{2}, \forall i \in \{1, \ldots, K\},$$

or equivalently:

$$T_i(n) \geqslant \frac{4}{25}\frac{a}{\Delta_i^2}, \forall i \in \{1, \ldots, K\}. \tag{7.14}$$

Second step. Firstly we prove by induction that

$$T_i(t) \leqslant \frac{36}{25}\frac{a}{\Delta_i^2} + 1, \forall i \neq i^*. \tag{7.15}$$

It is obviously true at time $t = 1$. Now assume that the formula is true at time $t-1$. If $I_t \neq i$ then $T_i(t) = T_i(t-1)$ and the formula still holds. On the other hand, if $I_t = i$, then in particular it means that $B_{i,T_i(t-1)} \geqslant B_{i^*,T_{i^*}(t-1)}$. Moreover, since we are on ξ, we have $B_{i^*,T_{i^*}(t-1)} \geqslant \mu^*$ and $B_{i,T_i(t-1)} \leqslant \mu_i + \frac{6}{5}\sqrt{\frac{a}{T_i(t-1)}}$. Thus, we have $\frac{6}{5}\sqrt{\frac{a}{T_i(t-1)}} \geqslant \Delta_i$. By using $T_i(t) = T_i(t-1) + 1$, we obtain (7.15).

Now we prove an other useful formula:

$$T_i(t) \geqslant \frac{4}{25}\min\left(\frac{a}{\Delta_i^2}, \frac{25}{36}(T_{i^*}(t)-1)\right), \forall i \neq i^*. \tag{7.16}$$

With the same inductive argument as the one to get equation (7.15), we only need to prove that this formula holds when $I_t = i^*$. By definition of the algorithm, and since we are on ξ, when $I_t = i^*$ we have for all i:

$$\mu^* + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(t-1)}} \geqslant \mu_i + \frac{4}{5}\sqrt{\frac{a}{T_i(t-1)}},$$

which implies

$$T_i(t-1) \geqslant \frac{16}{25}\frac{a}{\left(\Delta_i + \frac{6}{5}\sqrt{\frac{a}{T_{i^*}(t-1)}}\right)^2}.$$

We then obtain (7.16) by using $u+v \leqslant 2\max(u,v)$, $T_i(t) = T_i(t-1)$ and $T_{i^*}(t-1) = T_{i^*}(t)-1$.
Third step. Recall that we want to prove equation (7.14). From (7.16), we only have to show that

$$\frac{25}{36}(T_{i^*}(n)-1) \geqslant \frac{a}{\Delta_{i^*}^2},$$

where we recall that Δ_{i^*} is the minimal gap $\Delta_{i^*} = \min_{i\neq i^*}\Delta_i$. Using equation (7.15) we obtain:

$$T_{i^*}(n) - 1 = n - 1 - \sum_{i\neq i^*} T_i(n) \geqslant n - K - \frac{36}{25}a\sum_{i\neq i^*}\Delta_i^{-2} \geqslant \frac{36}{25}a\Delta_{i^*}^{-2},$$

where the last inequality uses $\frac{36}{25}H_1 a \leqslant n - K$. This concludes the proof.

8.3. Lower bound for UCB-E.

THEOREM 7.5. *If $\nu_2, \dots, \nu_K$ are Dirac distributions concentrated at $\frac{1}{2}$ and if ν_1 is the Bernoulli distribution of parameter $3/4$, the UCB-E algorithm satisfies $4\mathbb{E}r_n = e_n \geqslant 4^{-(4a+1)}$.*

PROOF. Consider the event $\mathcal{E}$ on which the reward obtained from the first $m = \lceil 4a \rceil$ draws of arm 1 are equal to zero. On this event of probability 4^{-m}, UCB-E will not draw arm 1 more than m times. Indeed, if it is drawn m times, it will not be drawn another time since $B_{1,m} \leqslant \frac{1}{2} < B_{2,s}$ for any s. On the event $\mathcal{E}$, we have $J_n \neq 1$. □

8.4. Application of Hoeffding's maximal inequality in the proof of Theorem 7.4. Let $i \in \{2, \dots, L\}$ and $j \in \{1, \dots, L\}$. First note that, by definition of ν' and since $i \neq 1$,

$$\mathbb{E}_{\nu'}\widehat{\mathrm{KL}}_{i,t}(\nu_i, \nu_j) = t\ \mathrm{KL}(\nu_i, \nu_j).$$

Since $\nu_i = \mathrm{Ber}(\mu_i)$ and $\nu_j = \mathrm{Ber}(\mu_j)$, with $\mu_i, \mu_j \in [p, 1-p]$, we have

$$\left|\log\left(\frac{d\nu_i(X_{i,t})}{d\nu_j(X_{i,t})}\right)\right| \leqslant \log(p^{-1}).$$

From Hoeffding's maximal inequality, see e.g. [Cesa-Bianchi and Lugosi, 2006, Section A.1.3], we have to bound almost surely the quantity, with $\mathbb{P}_{\nu'}$-probability at least $1 - \frac{1}{2L^2}$, we have for all $t \in \{1, \dots, n\}$,

$$\widehat{\mathrm{KL}}_{i,t}(\nu_i, \nu_j) - t\ \mathrm{KL}(\nu_i, \nu_j) \leqslant 2\log(p^{-1})\sqrt{\frac{\log(L^2)n}{2}}.$$

Similarly, with $\mathbb{P}_{\nu'}$-probability at least $1 - \frac{1}{2L^2}$, we have for all $t \in \{1, \dots, n\}$,

$$\widehat{\mathrm{KL}}_{1,t}(\nu_L, \nu_j) - t\ \mathrm{KL}(\nu_L, \nu_j) \leqslant 2\log(p^{-1})\sqrt{\frac{\log(L^2)n}{2}}.$$

A simple union bound argument then gives $\mathbb{P}_{\nu'}(C_n) \geqslant 1/2$.

Part 2

Clustering Foundations

CHAPTER 8

Nearest Neighbor Clustering: A Baseline Method for Consistent Clustering with Arbitrary Objective Functions

Clustering is often formulated as a discrete optimization problem. The objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the underlying space. We argue that the discrete optimization approach usually does not achieve this goal, and instead can lead to inconsistency. We construct examples which provably have this behavior. As in the case of supervised learning, the cure is to restrict the size of the function classes under consideration. For appropriate "small" function classes we can prove very general consistency theorems for clustering optimization schemes. As one particular algorithm for clustering with a restricted function space we introduce "nearest neighbor clustering". Similar to the k-nearest neighbor classifier in supervised learning, this algorithm can be seen as a general baseline algorithm to minimize arbitrary clustering objective functions. We prove that it is statistically consistent for all commonly used clustering objective functions.

Contents

This chapter is a joint work with Ulrike Von Luxburg. It is based on the paper Bubeck and von Luxburg [2009] published in the Journal of Machine Learning Research. A previous version, von Luxburg et al. [2008], appeared in Advances in Neural Information Processing Systems 21.

1. Introduction

Clustering is the problem of discovering "meaningful" groups in given data. In practice, the most common approach to clustering is to define a clustering quality function Q_n, and then construct an algorithm which is able to minimize (or maximize) Q_n. There exists a huge variety of clustering quality functions: the K-means objective function based on the distance of the data points to the cluster centers, graph cut based objective functions such as ratio cut or normalized cut, or various criteria based on some function of the within- and between-cluster similarities. Once a particular clustering quality function Q_n has been selected, the objective of clustering is stated as a discrete optimization problem. Given a data set $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and a clustering quality function Q_n, the ideal clustering algorithm should take into account all possible partitions of the data set and output the one that minimizes Q_n. The implicit understanding is that the "best" clustering can be any partition out of the set of all possible partitions of the data set. The practical challenge is then to construct an algorithm which is able to explicitly compute this "best" clustering by solving an optimization problem. We will call this approach the "discrete optimization approach to clustering".

Now let us look at clustering from the perspective of statistical learning theory. Here we assume that the finite data set has been sampled from an underlying data space $\mathcal{X}$ according to some probability measure $\mathbb{P}$. The ultimate goal in this setting is not to discover the best possible partition of the data set $\mathcal{X}_n$, but to learn the "true clustering" of the underlying space $\mathcal{X}$. While it is not obvious how this "true clustering" should be defined in a general setting [cf. von Luxburg and Ben-David, 2005], in an approach based on quality functions this is straightforward. We choose a clustering quality function Q on the set of partitions of the entire data space $\mathcal{X}$, and define the true clustering f^* to be the partition of $\mathcal{X}$ which minimizes Q. In a finite sample setting, the goal is now to approximate this true clustering as well as possible. To this end, we define an empirical quality function Q_n which can be evaluated based on the finite sample only, and construct the empirical clustering f_n as the minimizer of Q_n. In this setting, a very important property of a clustering algorithm is consistency: we require that $Q(f_n)$ converges to $Q(f^*)$ when $n \to \infty$. This strongly reminds of the standard approach in supervised classification, the empirical risk minimization approach. For this approach, the most important insight of statistical learning theory is that in order to be consistent, learning algorithms have to choose their functions from some "small" function space only. There are many ways how the size of a function space can be quantified. One of the easiest ways is to use shattering coefficients $s(\mathcal{F}, n)$ (see Section 2 for details). A typical result in statistical learning theory is that a necessary condition for consistency is $\mathbb{E} \log s(\mathcal{F}, n)/n \to 0$ (cf. Theorem 2.3 in Vapnik, 1995, Section 12.4 of Devroye et al., 1996). That is, the "number of functions" $s(\mathcal{F}, n)$ in $\mathcal{F}$ must not grow exponentially in n, otherwise one cannot guarantee for consistency.

Stated like this, it becomes apparent that the two viewpoints described above are not compatible with each other. While the discrete optimization approach on any given sample attempts to find the best of all (exponentially many) partitions, statistical learning theory suggests to restrict the set of candidate partitions to have sub-exponential size. So from the statistical learning theory perspective, an algorithm which is considered ideal in the discrete optimization setting will not produce partitions which converge to the true clustering of the data space.

In practice, for most clustering objective functions and many data sets the discrete optimization approach cannot be performed perfectly as the corresponding optimization problem is NP hard. Instead, people resort to heuristics and accept suboptimal solutions. One approach is to use local optimization procedures potentially ending in local minima only. This is what happens in the K-means algorithm: even though the K-means problem for fixed K and fixed dimension is not NP hard, it is still too hard for being solved globally in practice. Another approach is to construct a relaxation of the original problem which can be solved efficiently (spectral clustering is an example for this). For such heuristics, in general one cannot guarantee how close the heuristic solution is to the finite sample optimum. This situation is clearly unsatisfactory: in general, we neither have guarantees on the finite sample behavior of the algorithm, nor on its statistical consistency in the limit.

The following alternative approach looks much more promising. Instead of attempting to solve the discrete optimization problem over the set of all partitions, and then resorting to relaxations due to the hardness of this problem, we turn the tables. Directly from the outset, we only consider candidate partitions in some restricted class $\mathcal{F}_n$ containing only polynomially many functions. Then the discrete optimization problem of minimizing Q_n over $\mathcal{F}_n$ is not NP hard—formally it can be solved in polynomially many steps by trying all candidates in $\mathcal{F}_n$. From a theoretical point of view this approach has the advantage that the resulting clustering algorithm has the potential of being consistent. In addition, this approach also has advantages in practice: rather than dealing with uncontrolled relaxations of the original problem, we restrict the function class to some small subset $\mathcal{F}_n$ of "reasonable" partitions. Within this subset, we then have complete control over the solution of the optimization problem and can find the global optimum. Put another way, one can also interpret this approach as some controlled way to approximate a solution of the NP hard optimization problem on the finite sample, with the positive side effect of obeying the rules of statistical learning theory.

This is the approach we want to describe in this chapter. In Section 2 we will first construct an example which demonstrates the inconsistency in the discrete optimization approach. Then we will state a general theorem which gives sufficient conditions for clustering optimization schemes to be consistent. We will see that the key point is to control the size of the function classes the clustering are selected from. In Section 3 we will then introduce an algorithm which is able to work with such a restricted function class. This algorithm is called nearest neighbor clustering, and in some sense it can be seen as a clustering-analogue to the well-known nearest neighbor classifier for classification. We prove that nearest neighbor clustering is consistent under minimal assumptions on the clustering quality functions Q_n and Q. Then we will apply nearest neighbor clustering to a large variety of clustering objective functions, such as the K-means objective function, normalized cut and ratio cut, the modularity objective function, or functions based on within-between cluster similarity ratios. For all these functions we will verify the consistency of nearest neighbor clustering in Section 4. Discussion of our results, also in the context of the related literature, can be found in Sections 5 and 6. The proofs of all our results are deferred to Section 7, as some of them are rather technical.

2. General (In)Consistency Results

In the rest of this chapter, we consider a space $\mathcal{X}$ which is endowed with a probability measure $\mathbb{P}$. The task is to construct a clustering $f : \mathcal{X} \to \{1, \ldots, K\}$ on this space, where K denotes the number of clusters to construct. We denote the space of all $\mathbb{P}$-measurable functions from $\mathcal{X}$ to $\{1, ..., K\}$ by $\mathcal{H}$. Let $Q : \mathcal{H} \to \mathbb{R}^+$ denote a clustering quality function: for each clustering, it tells us "how good" a given clustering is. This quality function will usually depend on the

probability measure $\mathbb{P}$. An optimal clustering, according to this objective function, is a clustering f^* which satisfies

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} Q(f).$$

where $\mathcal{F} \subseteq \mathcal{H}$ is a fixed set of candidate clusterings. Now assume that $\mathbb{P}$ is unknown, but that we are given a finite sample $X_1, ..., X_n \in \mathcal{X}$ which has been drawn i.i.d according to $\mathbb{P}$. Our goal is to use this sample to construct a clustering f_n which "approximates" an optimal clustering f^*. To this end, assume that $Q_n : \mathcal{H} \to \mathbb{R}^+$ is an estimator of Q which can be computed based on the finite sample only (that is, it does not involve any function evaluations $f(x)$ for $x \notin \{X_1, ..., X_n\}$). We then consider the clustering

$$f_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q_n(f).$$

Here, $\mathcal{F}_n$ is a subset of $\mathcal{H}$, which might or might not be different from $\mathcal{F}$. The general question we are concerned with in this chaper is the question of consistency: under which conditions do we know that $Q(f_n) \to Q(f^*)$?

Note that to avoid technical overload we will assume throughout this chapter that all the minima (as in the definitions of f^* and f_n) exist and can be attained. If this is not the case, one can always go over to statements about functions which are ε-close to the corresponding infimum. We also will not discuss issues of measurability in this chapter (readers interested in measurability issues for empirical processes are referred to Section 1 of van der Vaart and Wellner, 1996).

2.1. Inconsistency example. In the introduction we suggested that as in the supervised case, the size of the function class $\mathcal{F}_n$ might be the key to consistency of clustering. In particular, we argued that optimizing over the space of all measurable functions might lead to inconsistency. First of all, we would like to prove this statement by providing an example. This example will show that if we optimize a clustering objective function over a too large class of functions, the resulting clusterings are not consistent.

EXAMPLE 8.1 (Inconsistency in general). *As data space we choose $\mathcal{X} = [0, 1] \cup [2, 3]$, and as probability measure $\mathbb{P}$ we simply use the normalized Lebesgue measure λ on $\mathcal{X}$. We define the following similarity function between points in $\mathcal{X}$:*

$$s(x, y) = \begin{cases} 1 & \text{if } x \in [0, 1], y \in [0, 1] \\ 1 & \text{if } x \in [2, 3],\ y \in [2, 3] \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we consider the case where we want to construct $K = 2$ clusters called C_1 and C_2. Given a clustering function $f : \mathcal{X} \to \{0, 1\}$ we call the clusters $C_1 := \{x \in \mathcal{X} \mid f(x) = 0\}$ and $C_2 := \{x \in \mathcal{X} \mid f(x) = 1\}$. As clustering quality function Q we use the between-cluster similarity (equivalent to cut*, see Section* 4.2 *for details):*

$$Q(f) = \int_{x \in C_1} \int_{y \in C_2} s(X, Y)\, d\mathbb{P}(X)\, d\mathbb{P}(Y).$$

As an estimator of Q we will use the function Q_n where the integrals are replaced by sums over the data points:

$$Q_n(f) = \frac{1}{n(n-1)} \sum_{i \in C_1} \sum_{j \in C_2} s(X_i, X_j).$$

As set $\mathcal{F}$ we choose the set of all measurable partitions on $\mathcal{X}$ (note that the same example also holds true when we only look at the set $\mathcal{F}$ of measurable partitions such that both clusters have a minimal mass ε for some $\varepsilon > 0$). For all $n \in \mathbb{N}$ we set $\mathcal{F}_n = \mathcal{F}$. Let $X_1, ..., X_n \in \mathcal{X}$ be our training data. Now define the functions

$$f^*(x) = \begin{cases} 0 & \text{if } x \in [0,1] \\ 1 & \text{if } x \in [2,3] \end{cases} \qquad \text{and} \qquad f_n(x) = \begin{cases} 0 & \text{if } x \in \{X_1, ..., X_n\} \cap [0,1] \\ 1 & \text{if } x \in [2,3] \\ 0 & \text{if } x \in [0, 0.5] \setminus \{X_1, ..., X_n\} \\ 1 & \text{if } x \in [0.5, 1] \setminus \{X_1, ..., X_n\} \end{cases}.$$

It is obvious that $Q(f^*) = 0$ and $Q_n(f_n) = 0$. As both Q and Q_n are non-negative, we can conclude $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} Q(f)$ and $f_n \in \operatorname{argmin}_{f \in \mathcal{F}} Q_n(f)$. It is also straightforward to compute $Q(f_n) = 1/16$ (independently of n). Hence, we have inconsistency: $1/16 = Q(f_n) \not\to Q(f^*) = 0$.

Note that the example is set up in a rather natural way. The data space contains two perfect clusters ($[0,1]$ and $[2,3]$) which are separated by a large margin. The similarity function is the ideal similarity function for this case, giving similarity 1 to points which are in the same cluster, and similarity 0 to points in different clusters. The function f^* is the correct clustering. The empirical clustering f_n, if restricted to the data points, reflects the correct clustering. It is just the "extension" of the empirical clustering to non-training points which leads to the inconsistency of f_n. Intuitively, the reason why this can happen is clear: the function space $\mathcal{F}$ does not exclude the unsuitable extension chosen in the example, the function overfits. This can happen because the function class is too large.

2.2. Main Result. Now we would like to present our first main theorem. It shows that if f_n is only picked out of a "small" function class $\mathcal{F}_n$, then we can guarantee consistency of clustering. Before stating the theorem we would like to recall the definition of the shattering coefficient in a K-class setting. For a function class $\mathcal{F} : \mathcal{X} \to \{1, \dots, K\}$ the shattering coefficient of size n is defined as

$$s(\mathcal{F}, n) = \max_{x_1, \dots, x_n \in \mathcal{X}} |\{(f(x_1), \dots, f(x_n)) \mid f \in \mathcal{F}\}|.$$

To state our theorem, we will also require a pseudo-distance d between functions. A pseudo-distance is a dissimilarity function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ which is symmetric, satisfies the triangle inequality and the condition $f = g \implies d(f,g) = 0$, but not necessarily the condition $d(f,g) = 0 \implies f = g$. For distances between sets of functions we use the standard convention $d(\mathcal{F}, \mathcal{G}) = \inf_{f \in \mathcal{F}, g \in \mathcal{G}} d(f,g)$. Our theorem is as follows:

THEOREM 8.1 (Consistency of a clustering optimizing scheme). *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables which have been drawn i.i.d. according to some probability measure $\mathbb{P}$ on some set $\mathcal{X}$. Let $\mathcal{F}_n := \mathcal{F}_n(X_1, \dots, X_n) \subset \mathcal{H}$ be a sequence of function spaces, and $\mathcal{F} \subset \mathcal{H}$. Let $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a pseudo-distance defined on $\mathcal{H}$. Let $Q : \mathcal{H} \to \mathbb{R}^+$ be a clustering quality function, and $Q_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function which can be computed based on the finite sample only. Finally let*

$$\widetilde{\mathcal{F}_n} := \bigcup_{X_1, \dots, X_n \in \mathbb{R}^d} \mathcal{F}_n.$$

Define the true and the empirical clusterings as

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} Q(f),$$

$$f_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q_n(f).$$

Assume that the following conditions are satisfied:

(1) $Q_n(f)$ *is a consistent estimator of* $Q(f)$ *which converges sufficiently fast for all* $f \in \widetilde{\mathcal{F}_n}$:

$$\forall \varepsilon > 0, \quad s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| > \varepsilon) \to 0,$$

(2) $\mathcal{F}_n$ *approximates* $\mathcal{F}$ *in the following sense:*

$$(i) \quad \forall f \in \mathcal{F}, d(f, \mathcal{F}_n) \to 0 \text{ in probability,}$$
$$(ii) \quad \mathbb{P}(f_n \notin \mathcal{F}) \to 0.$$

(3) Q *is uniformly continuous with respect to the pseudo-distance* d *between* $\mathcal{F}$ *and* $\widetilde{\mathcal{F}_n}$*:*

$$\forall \varepsilon > 0 \; \exists \delta(\varepsilon) > 0 \text{ such that } \forall f \in \mathcal{F} \; \forall g \in \widetilde{\mathcal{F}_n} : \; d(f,g) \leqslant \delta(\varepsilon) \Rightarrow |Q(f) - Q(g)| \leqslant \varepsilon.$$

Then the optimization scheme is weakly consistent, that is $Q(f_n) \to Q(f^*)$ *in probability.*

This theorem states sufficient conditions for consistent clustering schemes. In the context of the standard statistical learning theory, the three conditions in the theorem are rather natural. The first condition mainly takes care of the estimation error. Implicitly, it restricts the size of the function class $\mathcal{F}_n$ by incorporating the shattering coefficient. We decided to state condition 1 in this rather abstract way to make the theorem as general as possible. We will see later how it can be used in concrete applications. Of course, there are many more ways to specify the size of function classes, and many of them might lead to better bounds in the end. However, in this chapter we are not so much concerned with obtaining the sharpest bounds, but we want to demonstrate the general concept (as the reader can see in appendix, the proofs are already long enough using simple shattering numbers). The second condition in the theorem takes care of the approximation error. Intuitively it is clear that if we want to approximate solutions in $\mathcal{F}$, eventually $\mathcal{F}_n$ needs to be "close" to $\mathcal{F}$. The third condition establishes a relation between the quality function Q and the distance function d: if two clusterings f and g are close with respect to d, then their quality values $Q(f)$ and $Q(g)$ are close, too. We need this property to be able to conclude from "closeness" as in Condition 2 to "closeness" of the clustering quality values.

Finally, we would like to point out a few technical treats. First of all, note that the function class $\mathcal{F}_n$ is allowed to be data dependent. Secondly, as opposed to most results in empirical risk minimization we do not assume that Q_n is an unbiased estimator of Q (that is, we allow $\mathbb{E}Q_n \neq Q$), nor does Q need to be "an expectation" (that is, of the form $Q(f) = \mathbb{E}(\Omega(f, X))$ for some Ω). Both facts make the proof more technical, as many of the standard tools (symmetrization, concentration inequalities) become harder to apply. However, this is necessary since in the context of clustering biased estimators pop up all over the place. We will see that many of the popular clustering objective functions lead to biased estimators.

3. Nearest Neighbor Clustering—General Theory

The theorem presented in the last section shows sufficient conditions under which clustering can be performed consistently. Now we want to present a generic algorithm which can be used to minimize arbitrary clustering objective functions. With help of Theorem 8.1 we can then prove the consistency of its results for a large variety of clustering objective functions.

We have seen that the key to obtain consistent clustering schemes is to work with an appropriate function class. But of course, given quality functions Q and Q_n, the question is how such a function space can be constructed in practice. Essentially, three requirements have to be satisfied:

- The function space $\mathcal{F}_n$ has to be "small". Ideally, it should only contain polynomially many functions.
- The function space $\mathcal{F}_n$ should be "rich enough". In the limit $n \to \infty$, we would like to be able to approximate any (reasonable) measurable function.
- We need to be able to solve the optimization problem $\operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$. This sounds trivial at first glance, but in practice is far from easy.

One rather straightforward way to achieve all requirements is to use a function space of piecewise constant functions. Given a partitioning of the data space in small cells, we only look at clusterings which are constant on each cell (that is, the clustering never splits a cell). If we make sure that the number of cells is only of the order $\log(n)$, then we know that the number of clusterings is at most $K^{\log(n)} = n^{\log(K)}$, which is polynomial in n. In the following we will introduce a data-dependent random partition of the space which turns out to be very convenient.

3.1. Nearest Neighbor Clustering—The Algorithm. We will construct a function class $\mathcal{F}_n$ as follows. Given a finite sample $X_1, \ldots, X_n \in \mathbb{R}^d$, the number K of clusters to construct, and a number $m \in \mathbb{N}$ with $K \leqslant m \ll n$, randomly pick a subset of m "seed points" $X_{s_1}, \ldots, X_{s_m}$. Assign all other data points to their closest seed points, that is for all $j = 1, \ldots, m$ define the set Z_j as the subset of data points whose nearest seed point is X_{s_j}. In other words, the sets $Z_1, \ldots, Z_m$ are the Voronoi cells induced by the seeds $X_{s_1}, \ldots, X_{s_m}$. Then consider all partitions of $\mathcal{X}_n$ which are constant on all the sets $Z_1, ..., Z_m$. More formally, for given seeds we define the set $\mathcal{F}_n$ as the set of all functions

$$\mathcal{F}_n := \{f : \mathcal{X} \to \{1, \ldots, K\} \mid \forall\, j = 1, \ldots, m : \forall z, z' \in Z_j : f(z) = f(z')\}.$$

Obviously, the function class $\mathcal{F}_n$ contains K^m functions, which is polynomial in n if the number m of seeds satisfies $m \in O(\log n)$. Given $\mathcal{F}_n$, the most simple polynomial-time optimization algorithm is then to evaluate $Q_n(f)$ for all $f \in \mathcal{F}_n$ and choose the solution $f_n = \operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$. We call the resulting clustering the *nearest neighbor clustering* and denote it by $\mathrm{NNC}(Q_n)$. The entire algorithm is summarized in Figure 1. We have already published results on the empiri-

Nearest Neighbor Clustering NNC(Q_n), naive implementation

Parameters: number K of clusters to construct, number $m \in \mathbb{N}$ of seed points to use (with $K \leqslant m \ll n$), clustering quality function Q_n

Input: data set $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, distances $d_{ij} = d(X_i, X_j)$

- Subsample m seed points from the data points, without replacement.
- Build the Voronoi decomposition $Z_1, \ldots, Z_m$ of $\mathcal{X}_n$ based on the distances d_{ij} using the seed points as centers
- Define $\mathcal{F}_n := \{f : \mathcal{X}_n \to \{1, \ldots, K\} \mid f$ constant on all cells $Z_j\}$
- For all $f \in \mathcal{F}_n$ evaluate $Q_n(f)$.

Output: $f_n := \operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$

Figure 1: Nearest neighbor clustering for a general clustering objective function Q_n.

cal performance of the algorithm in von Luxburg et al. [2008], and more results can be found in

Section 3 of Jegelka [2007]. We have found that on finite samples, the algorithm performs surprisingly well in terms of quality function: using $m = \log n$ seed points, the objective function values obtained at the solutions are comparable to these of K-means or spectral clustering, respectively. Moreover, there exist efficient ways to compute f_n using branch and bound methods. Using these methods, the running time of nearest neighbor clustering using $m = \log n$ seeds is roughly comparable to the one of the other clustering algorithms. See von Luxburg et al. [2008] and Jegelka [2007] for details on the experimental results.

3.2. Consistency of Nearest Neighbor Clustering (General Statement). Now we want to prove that nearest neighbor clustering is consistent. We will see that even though we can rely on Theorem 8.1, the consistency proof for nearest neighbor clustering does not come for free. Let $f : \mathcal{X} \to \{1, \dots, K\}$ be a clustering function. In the following, we will often use the notation f_k for the indicator function of the k-th cluster:

$$f_k(x) := \mathbb{1}_{f(x)=k}.$$

This is a slight abuse of notation, as we already reserved the notation f_n for the minimizer of the empirical quality function. However, from the context it will always be clear whether we will refer to f_n or f_k, respectively, as we will not mix up the letters n (for the sample size) and k (a cluster index).

As distance function between two clusterings we use the 0-1-loss

$$d(f, g) := \mathbb{P}(f(X) \neq g(X) | X_1, \dots, X_n).$$

Here the conditioning is needed for the cases where the functions f or g are data dependent. Note that in clustering, people often consider a variant of this distance which is independent with respect to the choice of labels, that is they choose $\widetilde{d}(f, g) := \min_\pi P(f(X) \neq \pi(g(X)) | X_1, \dots, X_n)$, where π runs over all permutations of the set $\{1, \dots, K\}$. However, we will see that for our purposes it does not hurt to use the overly sensitive 0-1 distance instead. The main reason is that at the end of the day, we only want to compare functions based on their quality values, which do not change under label permutations. In general, the theorems and proofs could also be written in terms of $\widetilde{d}$. For better readability, we decided to stick to the standard 0-1 distance, though.

We will see below that in many cases, even in the limit case one would like to use a function space $\mathcal{F}$ which is a proper subset of $\mathcal{H}$. For example, one could only be interested in clusterings where all clusters have a certain minimal size, or where the functions satisfy certain regularity constraints. In order to be able to deal with such general function spaces, we will introduce a tool to restrict function classes to functions satisfying certain conditions. To this end, let

$$\Phi : \mathcal{H} \to \mathbb{R}^+$$

be a functional which quantifies certain aspects of a clustering. In most cases, we will use functionals Φ which operate on the individual cluster indicator functions f_k. For example, $\Phi(f_k)$ could measure the size of cluster k, or the smoothness of the cluster boundary. The function class $\mathcal{F}$ will then be defined as

$$\mathcal{F} = \{f \in \mathcal{H} \mid \Phi(f_k) > a \text{ for all } k = 1, \dots, K\},$$

where $a \geqslant 0$ is a constant. In general, the functional Φ can be used to encode our intuition about "what a cluster is". Note that this setup also includes the general case of $\mathcal{F} = \mathcal{H}$, that is the case where we do not want to make any further restrictions on $\mathcal{F}$, for example by setting $\Phi(f_k) \equiv 1$, $a \equiv 0$. As it is the case for Q, we will usually not be able to compute Φ on a finite sample only. Hence we also introduce an empirical counterpart Φ_n which will be used in the finite sample case.

The following theorem will state sufficient conditions for the consistency of nearest neighbor clustering. For simplicity we state the theorem for the case $\mathcal{X} = \mathbb{R}^d$, but the proofs can also be carried over to more general spaces. Also, note that we only state the theorem for the case $d \geqslant 2$; in case $d = 1$ the theorem holds as well, but the formulas look a bit different.

THEOREM 8.2 (Consistency of nearest neighbor clustering). *Let $\mathcal{X} = \mathbb{R}^d$, $d \geqslant 2$, $Q : \mathcal{H} \to \mathbb{R}^+$ be a clustering quality function, and $Q_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function which can be computed based on the finite sample only. Similarly, let $\Phi : \mathcal{H} \to \mathbb{R}^+$, and $\Phi_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function. Let $a > 0$ and $(a_n)_{n\in\mathbb{N}}$ be such that $a_n > a$ and $a_n \to a$. Let $m = m(n) \leqslant n \in \mathbb{N}$. Finally, denote $d(f, g)$ the 0-1-loss, and let $NN_m(x)$ be the nearest neighbor of x among $X_1, \ldots, X_m$ according to the Euclidean distance. Define the function spaces*

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f \text{ continuous a.e. and } \forall k \in \{1, \ldots, K\} \;\; \Phi(f_k) > a\}$$

$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1, \ldots, K\} \;\; \Phi_n(f_k) > a_n\}$$

$$\widetilde{\mathcal{F}_n} := \bigcup_{X_1,\ldots,X_n\in\mathbb{R}^d} \mathcal{F}_n$$

$$\widehat{\mathcal{F}_n} := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid \exists \text{ Voronoi partition of m cells: } f \text{ constant on all cells}\}.$$

Assume that the following conditions are satisfied:

(1) *$Q_n(f)$ is a consistent estimator of $Q(f)$ which converges sufficiently fast for all $f \in \widetilde{\mathcal{F}_n}$:*

$$\forall \varepsilon > 0, K^m(2n)^{(d+1)m^2} \sup_{f\in\widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| > \varepsilon) \to 0,$$

(2) *$\Phi_n(f_k)$ is a consistent estimator of $\Phi(f_k)$ which converges sufficiently fast for all $f \in \widehat{\mathcal{F}_n}$:*

$$\forall \varepsilon > 0, K^m(2n)^{(d+1)m^2} \sup_{f\in\widehat{\mathcal{F}_n}} \mathbb{P}(|\Phi_n(f_k) - \Phi(f_k)| > \varepsilon) \to 0,$$

(3) *Q is uniformly continuous with respect to the pseudo-distance $d(f, g)$ between $\mathcal{F}$ and $\widetilde{\mathcal{F}_n}$, as defined in Condition (3) of Theorem 8.1,*

(4) *$\Phi_k(f) := \Phi(f_k)$ is uniformly continuous with respect to the pseudo-distance $d(f, g)$ between $\mathcal{F}$ and $\widehat{\mathcal{F}_n}$, as defined in Condition (3) of Theorem 8.1,*

(5) *a_n decreases slowly enough to a:*

$$K^m(2n)^{(d+1)m^2} \sup_{g\in\widehat{\mathcal{F}_n},k} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \geqslant a_n - a) \to 0,$$

(6) *$m \to \infty$.*

Then nearest neighbor clustering based on m seed points using quality function Q_n is weakly consistent, that is for $f_n \in \operatorname{argmin}_{f\in\mathcal{F}_n} Q_n(f)$ and $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} Q(f)$ we have $Q(f_n) \to Q(f^*)$ in probability.

This theorem is still rather abstract, but pretty powerful. In the following we will demonstrate this by applying it to many concrete clustering objective functions. To define our objective functions, we will from now on adopt the convention $0/0 = 0$.

4. Nearest Neighbor Clustering with Popular Clustering Objective Functions

In this section we want to study the consistency of nearest neighbor clustering when applied to particular objective functions. For simplicity we assume in this section that $\mathcal{X} = \mathbb{R}^d$.

4.1. NNC Using the K-means Objective Function. The K-means objective function is the within-cluster sum of squared distances, called WSS for short. To define it properly, for a given clustering function $f : \mathbb{R}^d \to \{1, \ldots, K\}$ we introduce the following quantities:

$$\mathrm{WSS}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_{k,n}\|^2 \qquad \text{where}$$

$$c_{k,n} := \frac{1}{n_k} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i \qquad \text{and} \qquad n_k := \frac{1}{n} \sum_{i=1}^{k} f_k(X_i)$$

$$\mathrm{WSS}(f) := \mathbb{E} \sum_{k=1}^{K} f_k(X) \|X - c_k\|^2 \qquad \text{where} \qquad c_k := \frac{\mathbb{E} f_k(X) X}{\mathbb{E} f_k(X)}.$$

Here, WSS_n plays the role of Q_n and WSS the role of Q. Let us point out some important facts. First the empirical quality function is not an unbiased estimator of the true one, that is $\mathbb{E}\,\mathrm{WSS}_n \neq \mathrm{WSS}$ and $\mathbb{E} c_{k,n} \neq c_k$ (note that in the standard treatment of K-means this can be achieved, but not on arbitrary function classes, see below for some discussion). However, at least we have $\mathbb{E} n_k = \mathbb{E} f_k(X)$ and $\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i = \mathbb{E} f_k(X) X$. Moreover, one should remark that if we define $\mathrm{WSS}(\cdot, \mathbb{P}) := \mathrm{WSS}$ then $\mathrm{WSS}_n = \mathrm{WSS}(\cdot, \mathbb{P}_n)$ where $\mathbb{P}_n$ is the empirical distribution.

Secondly, our setup for proving the consistency of nearest neighbor clustering with the WSS objective function is considerably more complicated than proving the consistency of the global minimizer of the K-means algorithm (e.g., Pollard, 1981). The reason is that for the K-means algorithm one can use a very helpful equivalence which does not hold for nearest neighbor clustering. Namely, if one considers the minimizer of WSS_n in the space of *all possible partitions*, then one can see that the clustering constructed by this minimizer always builds a Voronoi partition with K cells; the same holds in the limit case. In particular, given the cluster centers $c_{k,n}$ one can reconstruct the whole clustering by assigning each data point to the closest cluster center. As a consequence, to prove the convergence of K-means algorithms one usually studies the convergence of the empirical cluster centers $c_{k,n}$ to the true centers c_k. However, in our case this whole chain of arguments breaks down. The reason is that the clusters chosen by nearest neighbor clustering *from the set* $\mathcal{F}_n$ are not necessarily Voronoi cells, they do not even need to be convex (all clusters are composed by small Voronoi cells, but the union of "small" Voronoi cells is not a "large" Voronoi cell). Also, it is not the case that each data point is assigned to the cluster corresponding to the closest cluster center. It may very well happen that a point x belongs to cluster C_i, but is closer to the center of another cluster C_j than to the center of its own cluster C_i. Consequently, we cannot reconstruct the nearest neighbor clustering from the centers of the clusters. This means that we cannot go over to the convergence of centers, which makes our proof considerably more involved than the one of the standard K-means case.

Due to these technical problems, it will be of advantage to only consider clusters which have a certain minimal size (otherwise, the cluster quality function WSS is not uniformly continuous). To achieve this, we use the functionals

$$\Phi_{\mathrm{WSS}}(f_k) := \mathbb{E} f_k(X), \qquad \Phi_{\mathrm{WSS}_n}(f_k) := n_k(f).$$

and will only consider clusterings where $\Phi(f_k) \geqslant a > 0$. In practice, this can be interpreted as a simple means to avoid empty clusters. The constant a can be chosen so small that its only effect is to make sure that each cluster contains at least one data point. The corresponding function spaces

are

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1, \dots, K\} \mid f \text{ continuous a.e. and } \forall k \in \{1, \dots, K\} \;\; \Phi_{\mathrm{WSS}}(f_k) > a\}$$
$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1, \dots, K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1, \dots, K\} \;\; \Phi_{\mathrm{WSS}_n}(f_k) > a_n\}$$

Moreover, for technical convenience we restrict our attention to probability measures which have a bounded support inside some large ball, that is which satisfy $\operatorname{supp} \mathbb{P} \subset B(0, A)$ for some constant $A > 0$. It is likely that our results also hold in the general case, but the proof would get even more complicated. With the notation of Theorem 8.2 we have:

THEOREM 8.3 (Consistency of NNC(WSS)). *Assume that $a_n > a, a_n \to a, m \to \infty$ and*

$$\frac{m^2 \log n}{n(a - a_n)^2} \to 0.$$

Then for all probability measures on $\mathbb{R}^d$ with bounded support, nearest neighbor clustering with WSS *is consistent, that is if $n \to \infty$ then $\mathrm{WSS}(f_n) \to \mathrm{WSS}(f^*)$ in probability.*

This theorem looks very nice and simple. The conditions on a_n and m are easily satisfied as soon as these quantities do not converge too fast. For example, if we define

$$a_n = a + \frac{1}{\log n} \quad \text{and} \quad m = \log n$$

then

$$\frac{m^2 \log n}{n(a_n - a)^2} = \frac{(\log n)^5}{n} \to 0.$$

Moreover, it is straightforward to see from the proofs that this theorem is still valid if we consider the objective functions WSS_n and WSS with $\|\cdot\|$ instead of $\|\cdot\|^2$. It also holds for any other norm, such as the p-norms $\|\cdot\|_p$. However, it does not necessarily hold for powers of norms (in this sense, the squared Euclidean norm is an exception). The proof shows that the most crucial property is

$$\|X_i - c_{k,n}\| - \|X_i - c_k\| \leqslant \text{const} \cdot \|c_{k,n} - c_k\|.$$

This is straightforward if the triangle inequality holds, but might not be possible for general powers of norms.

By looking more carefully at our proofs one can state the following rate of convergence:

THEOREM 8.4 (Convergence Rate for NNC(WSS)). *Assume that $\operatorname{supp} \mathbb{P} \subset B(0, A)$ for some constant $A > 0$ and that $n(a_n - a)^2 \to \infty$. Let $\varepsilon \leqslant 1$ and $a^* := \inf_k \mathbb{E} f_k^*(X) - a > 0$. Then there exists :*

$$N = N((a_n), a^*) \in \mathbb{N},$$
$$C_1 = C_1(a, a^*, \varepsilon, K, A) > 0, \qquad C_2 = C_2(a, a^*, \varepsilon, A, f^*, \mathbb{P}) > 0,$$
$$C_3 = C_3(a, d, \varepsilon, K, A) > 0, \qquad C_4 = C_4(a, d, A) > 0$$

such that for $n \geqslant N$ the following holds true:

$$\mathbb{P}(|\,\mathrm{WSS}(f_n) - \mathrm{WSS}(f^*)| \geqslant \varepsilon)$$
$$\leqslant C_1 e^{-C_2 m} + K^{m+1}(2n)^{(d+1)m^2} \left(C_3 e^{-C_4 \varepsilon^2 n} + 8K e^{-\frac{n(a_n - a)^2}{8}} \right).$$

At first glance, it seems very tempting to try to use the Borel-Cantelli lemma to transform the weak consistency into strong consistency. However, we do not have an explicit functional form

of dependency of C_2 on ε. The main reason is that in Lemma 8.3 (Appendix) the constant $b(\varepsilon)$ will be defined only implicitly. If one would like to prove strong consistency of nearest neighbor clustering with WSS one would have to get an explicit form of $b(\varepsilon)$ in Lemma 8.3.

For a general discussion relating the consistency result of NNC(WSS) in to the consistency results by Pollard [1981] and others see Section 5.

4.2. NNC Using Standard Graph-cut Based Objective Functions. In this section we want to look into the consistency of nearest neighbor clustering for graph based objective functions as they are used in spectral clustering (see von Luxburg, 2007 for details). Let $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a similarity function which is upper bounded by a constant C. The two main quantities we need to define graph-cut based objective functions are the cut and the volume. For a given cluster described by the cluster indicator function $f_k : \mathbb{R}^d \to \{0, 1\}$, we set

$$\mathrm{cut}(f_k) := \mathrm{cut}(f_k, \mathbb{P}) := \mathbb{E} f_k(X_1)(1 - f_k(X_2))s(X_1, X_2),$$

$$\mathrm{vol}(f_k) := \mathrm{vol}(f_k, \mathbb{P}) := \mathbb{E} f_k(X_1)s(X_1, X_2).$$

For $f \in \mathcal{H}$ we can then define the normalized cut and the ratio cut by

$$\mathrm{Ncut}(f) := \mathrm{Ncut}(f, \mathbb{P}) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)},$$

$$\mathrm{RatioCut}(f) := \mathrm{RatioCut}(f, \mathbb{P}) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathbb{E} f_k(X)}.$$

The empirical estimators of these objective functions will be $\mathrm{Ncut}(f, \mathbb{P}_n)$ and $\mathrm{RatioCut}(f, \mathbb{P}_n)$, in explicit formulas:

$$\mathrm{cut}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i)(1 - f(X_j))s(X_i, X_j),$$

$$\mathrm{vol}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i)s(X_i, X_j), \qquad n_k := \frac{1}{n} \sum_{i=1}^{k} f_k(X_i),$$

$$\mathrm{Ncut}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{\mathrm{vol}_n(f_k)}, \qquad \mathrm{RatioCut}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{n_k}.$$

Again we need to define how we will measure the size of the clusters. We will use

$$\Phi_{\mathrm{cut}}(f_k) := \mathrm{vol}(f_k), \qquad \Phi_{\mathrm{Ncut}}(f_k) := \mathrm{vol}(f_k), \qquad \Phi_{\mathrm{RatioCut}}(f_k) := \mathbb{E} f_k(X).$$

with the corresponding empirical quantities $\Phi_{\mathrm{cut}_n}, \Phi_{\mathrm{Ncut}_n}$ and $\Phi_{\mathrm{RatioCut}_n}$. Then, with the notations of Theorem 8.2, we have:

THEOREM 8.5 (Consistency of NNC(cut), NNC(Ncut) and NNC(RatioCut)). *Assume that the similarity function s is bounded by a constant $C > 0$, let $a_n > a$, $a_n \to a$, $m \to \infty$ and*

$$\frac{m^2 \log n}{n(a - a_n)^2} \to 0.$$

Then nearest neighbor clustering with cut, Ncut *and* RatioCut *is universally weakly consistent, that is for all probability measures, if $n \to \infty$ we have* $\mathrm{cut}(f_n) \to \mathrm{cut}(f^*)$, $\mathrm{Ncut}(f_n) \to \mathrm{Ncut}(f^*)$ *and* $\mathrm{RatioCut}(f_n) \to \mathrm{RatioCut}(f^*)$ *in probability.*

For these objective functions one can also state a rate of convergence. For sake of shortness we only state it for the normalized cut:

THEOREM 8.6 (Convergence Rate for $\mathrm{NNC}(\mathrm{Ncut})$). *Assume that the similarity function s is bounded by $C > 0$ and that $n(a_n - a)^2 \to \infty$. Let $\varepsilon \leqslant 1$ and $a^* := \inf_k vol(f_k^*) - a > 0$. Then there exist*

$$N = N((a_n), a^*) \in \mathbb{N},$$
$$C_1 = C_1(a, a^*, \varepsilon, K, C) > 0, \qquad C_2 = C_2(a, a^*, \varepsilon, C, K, f^*, \mathbb{P}) > 0,$$
$$C_3 = C_3(a, \varepsilon, K, C) > 0, \qquad C_4 = C_4(a, K, C) > 0.$$

such that for $n \geqslant N$ the following holds true:

$$\mathbb{P}(|\operatorname{Ncut}(f_n) - \operatorname{Ncut}(f^*)| \geqslant \varepsilon)$$
$$\leqslant C_1 e^{-C_2 m} + K^{m+1}(2n)^{(d+1)m^2}\left(C_3 e^{-C_4\varepsilon^2 n} + 8Ke^{-\frac{n(a_n-a)^2}{8}}\right).$$

4.3. NNC Using the Modularity Objective Function. A slightly different objective functions for graph clustering is the "modularity", which has been put forward by Newman [2006] for detecting communities in networks. In this chapter, the modularity is formulated as an objective function to find communities in a finite graph. However, as it is the case for Ncut or $\mathrm{RatioCut}$, the modularity cannot be directly minimized. Instead, a spectral relaxation has been developed to minimize the modularity, see Newman [2006] for details. Of course, the nearest neighbor clustering algorithm can also be used to minimize this objective function directly, without using a relaxation step. Using our own notation we define:

$$\operatorname{Mod}_n(f) =$$
$$\sum_{k=1}^{n} \frac{1}{n(n-1)} \sum_{i \neq j} f_k(X_i) f_k(X_j) \left(\frac{1}{(n-1)^2} \sum_{l, l \neq i} s(X_i, X_l) \sum_{l, l \neq j} s(X_j, X_l) - s(X_i, X_j) \right),$$
$$\operatorname{Mod}(f) =$$
$$\sum_{k=1}^{n} \int \int f_k(X) f_k(Y) \left(\int s(X, Z) d\mathbb{P}(Z) \int s(Y, Z) d\mathbb{P}(Z) - s(X, Y) \right) d\mathcal{P}(X, Y).$$

In the proof we will see that as the limit function $\mathrm{Mod}(\cdot)$ is uniformly continuous on $\mathcal{H}$, we do not need to quantify any function Φ or Φ_n to measure the volume of the clusters. The function classes are thus

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f \text{ continuous a.e.}\},$$
$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f(x) = f(NN_m(x))\}.$$

THEOREM 8.7 (Consistency of $\mathrm{NNC}(\mathrm{Mod})$). *Assume that $m \to \infty$ and*

$$\frac{m^2 \log n}{n} \to 0.$$

Then nearest neighbor clustering with Mod is universally weakly consistent: for all probability measures, if $n \to \infty$ then $\mathrm{Mod}(f_n) \to \mathrm{Mod}(f^*)$ in probability.

4.4. NNC Using Objective Function Based on the Ratio of Within-cluster and Between-cluster Similarity. Often, clustering algorithms try to minimize joint functions of the within-cluster similarity and the between cluster similarity. The most popular choice is the ratio of those two quantities, which is closely related to the criterion used in Fisher linear discriminant analysis.

Formally, the between-cluster similarity corresponds to the cut, and the within similarity of cluster k is given by

$$\mathrm{WS} := \mathbb{E} f(X_1) f(X_2) s(X_1, X_2).$$

Thus the ratio of between- and within-cluster similarity is given as

$$\mathrm{BWR}(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)}.$$

Again we use their empirical estimations:

$$\mathrm{WS}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i) f_k(X_j) s(X_i, X_j),$$

$$\mathrm{BWR}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)}.$$

To measure the size of the cluster we use

$$\Phi_{\mathrm{BWR}}(f_k) := \mathrm{WS}(f_k)$$

and its natural empirical counterpart. This leads to function spaces

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1, \dots, K\} \mid f \text{ continuous a.e. and } \forall k \in \{1, \dots, K\} \;\; \Phi_{\mathrm{BWR}}(f_k) > a\},$$

$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1, \dots, K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1, \dots, K\} \;\; \Phi_{\mathrm{BWR}_n}(f_k) > a_n\}.$$

THEOREM 8.8 (Consistency of $\mathrm{NNC}(\mathrm{BWR})$). *Assume that the similarity function s is bounded by a constant $C > 0$, let $a_n > a, a_n \to a, m \to \infty$ and*

$$\frac{m^2 \log n}{n(a - a_n)^2} \to 0.$$

Then nearest neighbor clustering with BWR *is universally weakly consistent, that is for all probability measure if $n \to \infty$ then* $\mathrm{BWR}(f_n) \to \mathrm{BWR}(f^*)$ *in probability.*

5. Relation to Previous Work

In this section we want to discuss our results in the light of the existing literature on consistent clusterings.

5.1. Standard Consistency Results for Center-based Algorithms. For a few clustering algorithms, consistency results are already known. The most well-known among them is the K-means algorithm. For this algorithm it has been first proved by Pollard [1981] that the global minimizer of the K-means objective function on a finite sample converges to the global minimizer on the underlying space.

First of all, we would like to point out that the consistency result by Pollard [1981] can easily be recovered using our theorems. Let us briefly recall the standard K-means setting. The objective function which K-means attempts to optimize is the function WSS, which we already encountered in the last sections. In the standard K-means setting the optimization problem is stated over the space of all measurable functions $\mathcal{H}$:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \mathrm{WSS}(f).$$

It is not difficult to prove that the solution f^* of this optimization problem always has a particular form. Namely, the solution f^* forms a Voronoi decomposition of the space, where the cluster

centers c_k are the centers of the Voronoi cells. Thus, we can rewrite the optimization problem above equivalently as

$$f^* = \operatorname*{argmin}_{f \in \mathcal{G}_K} \mathrm{WSS}(f)$$

where $\mathcal{G}_K$ denotes the set of all clusterings for which the clusters are Voronoi cells. The optimization problem for the finite sample case can be stated analogously:

$$f_n = \operatorname*{argmin}_{f \in \mathcal{G}_K} \mathrm{WSS}_n(f).$$

So in this particular case we can set $\mathcal{F}_n = \mathcal{F} = \mathcal{G}_K$. This means that even though the original optimization problem has been set up to optimize over the huge set $\mathcal{H}$, the optimization only needs to run over the small set $\mathcal{G}_K$. It is well known that the shattering coefficient of $\mathcal{G}_K$ is polynomial in n, namely it is bounded by $K^K n^{(d+1)K^2}$ (cf. Lemma 8.2). Moreover, the uniform continuity of WSS on $\mathcal{G}_K$ (Condition (3) of Theorem 8.2) can easily be verified if we assume that the probability distribution has compact support. As a consequence, using similar techniques as in the proofs of Theorem 8.3 we can prove that the global minimizer of the empirical K-means objective function WSS_n converges to the global minimizer of the true K-means objective function WSS. By this we recover the well-known result by Pollard [1981], under slightly different assumptions. In this sense, our Theorem 8.1 can be seen as a blueprint for obtaining Pollard-like results for more general objective functions and function spaces.

Are there any more advantages of Theorem 8.3 in the K-means setting? At first glance, our result in Theorem 8.3 looks similar to Pollard's result: the global minimizers of both objective functions converge to the true global minimizer. However, in practice there is one important difference. Note that as opposed to many vector quantization problems (cf. Garey et al., 1982), minimizing the K-means objective function is not NP-hard in n: the solution is always a Voronoi partition, there exist polynomially many Voronoi partitions of n points, and they can be enumerated in polynomial time (cf. Inaba et al., 1994). However, the size of the function class $\mathcal{G}_K$ is still so large that it would take too long to simply enumerate all its functions and select the best one. Namely, we will see in Lemma 8.2 that the number of Voronoi partitions of n points in $\mathbb{R}^d$ using K cells is bounded by $n^{(d+1)K}$, which is huge even for moderate d and K. As a work-around in practice one uses the well-known K-means *algorithm*, which is only able to find a *local* minimum of $\mathrm{WSS}_n(f)$. In contrast, nearest neighbor clustering works with a different function class which is much smaller than $\mathcal{G}_K$: it has only size $n^{\log K}$. On this smaller class we are still able to compute the *global* minimum of $\mathrm{WSS}_n(f)$. Consequently, our result in Theorem 8.3 is not only a theoretical statement about some abstract quantity as it is the case for Pollard's result, but it applies to the algorithm used in practice. While Pollard's result abstractly states that the global minimum (which cannot be computed efficiently) converges, our result implies that the result of nearest neighbor clustering does converge.

5.2. Consistency of Spectral Clustering. In the previous section we have seen in Theorems 8.5 and 8.6 that NNC is consistent for all the standard graph cut objective functions. Now we want to discuss these results in connection with the graph cut literature. It is well known that the discrete optimization problem of minimizing Ncut_n or $\mathrm{RatioCut}_n$ is an NP-hard problem, see Wagner and Wagner [1993]. However, approximate solutions of relaxed problems can be obtained by spectral clustering, see von Luxburg [2007] for a tutorial. Consistency results for spectral clustering algorithms have been proved in von Luxburg et al. [2008]. These results show that under certain conditions, the solutions computed by spectral clustering on finite samples converge to some kind of "limit solutions" based on the underlying distribution. In the light of the previous discussions,

this sounds plausible, as the space of solutions of spectral clustering is rather restricted: we only allow solutions which are eigenfunctions of certain integral operators. Thus, spectral clustering implicitly works with a small function class.

However, it is important to note that the convergence results of spectral clustering do not make any statement about the minimizers of Ncut (a similar discussion also holds for $\mathrm{RatioCut}$). The problem is that on any finite sample, spectral clustering only solves a relaxation of the original problem of minimizing Ncut_n. The Ncut_n-value of this solution can be arbitrarily far away from the minimal Ncut_n-value on this sample [Guattery and Miller, 1998], unless one makes certain assumptions which are not necessarily satisfied in a standard statistical setting (cf. Spielman and Teng, 1996, or Kannan et al., 2004). Thus the convergence statements for the results computed by the spectral clustering algorithm cannot be carried over to consistency results for the minimizers of Ncut. One knows that spectral clustering converges, but one does not have any guarantee about the Ncut-value of the solution. Here our results for nearest neighbor clustering present an improvement, as they directly refer to the minimizer of Ncut. While it is known that spectral clustering converges to "something", for the solutions computed by nearest neighbor clustering we know that they converge to the global minimizer of Ncut (or $\mathrm{RatioCut}$, respectively).

5.3. Consistency of Other Clustering Schemes. To the best of our knowledge, apart from results on center-based algorithms and spectral clustering, there are very few non-parametric clustering algorithms for which statistical consistency has been proved so far. The only other major class of algorithms for which consistency has been investigated is the class of linkage algorithms. While single linkage can be proved to be "fractionally consistent", that is it can at least discover sufficiently distinct high-density regions, both complete and average linkage are not consistent and can be misleading (cf. Hartigan, 1981, 1985). A more general method for hierarchical clustering used in Wong and Lane [1983] is statistically consistent, but essentially first estimates the density and then constructs density level sets based on this estimator.

Concerning parametric clustering algorithms, the standard setting is a model-based approach. One assumes that the underlying probability distribution has a certain parametric form (for example a mixture of Gaussians), and the goal is to estimate the parameters of the distribution from the sample. Estimating parameters in parametric models has been intensively investigated in statistics, in particular in the maximum likelihood framework and the Bayesian framework (for an overview how this can be done for clustering see Fraley and Raftery, 1998, or the book McLachlan and Peel, 2004). Numerous consistency results are known, but typically they require that the true underlying distribution indeed comes from the model class under consideration. For example, in a Bayesian setting one can show that in the large sample limit, the posterior distribution will concentrate around the true mixture parameters. However, if the model assumptions are not satisfied, counter-examples to consistency can be constructed. Moreover, the consistency results mentioned above are theoretic in the sense that the algorithm used in practice does not necessarily achieve them. Standard approaches to estimate mixture parameters are the EM algorithm (in a frequentist of MAP setting), or for example Markov Chain Monte Carlo sampling in a fully Bayesian approach. However, as it is the case for the K-means algorithm, these methods can get stuck in local optima, and no convergence towards the global optimum can be guaranteed. Another way to tackle model-based clustering problems is based on the minimum message length or minimum description length principle. The standard reference for MML approaches to learn mixtures is Figueiredo and Jain [2002], for a general overview on MDL see Grünwald [2007]. Consistency results for MML are quite similar to the ones for the Bayesian approach: if the true distribution indeed comes from the mixture class and the number of components is known, then consistency can be achieved.

For general results on consistency of MDL see Sections 16 and 17.11 in Grünwald [2007]. Often, MML/MDL approaches are interpreted as a particular way to work with small function classes, consisting of functions which can be described in a "compact" way. In this sense, this method can also be seen as a way of achieving "small" function classes.

5.4. Sublinear Time Algorithms Using Subsampling. Some algorithms related to our approach have been published in the theoretical computer science community, such as Indyk [1999], Mishra et al. [2001], or Czumaj and Sohler [2007]. The general idea is to use subsampling approaches to approximate clustering solutions, and to prove that these approximations are quite accurate. Given a sample of n points, one draws a subsample of $m \ll n$ points, applies some (approximate) clustering algorithm to the subsample, and then extends this clustering to the remaining points. Using techniques such as concentration inequalities, Chernoff bounds or Hoeffding bounds, one can then prove that the resulting clustering approximates the best clustering on the original point set.

While at first glance, this approach sounds very similar to our nearest neighbor clustering, note that the focus in these papers is quite a different one than ours. The authors do not aim for consistent clustering solutions (that is, solutions which are close to the "true clustering solution" of the underlying space), but they want to find algorithms to approximate the optimal clustering on a given finite sample in sublinear time. The sublinearity is achieved by the fact that already a very small subsample (say, $m = \log n$) is enough to achieve good approximation guarantees. However, our main point that it is important to control the size of the underlying function class, is not revealed in these papers. As the authors mainly deal with K-means type settings, they automatically work with polynomial function classes of center-based clusterings, and the issue of inconsistency does not arise. Moreover, subsampling is just one way of reducing the function class to a smaller size, there can be many others. In this sense, we believe that our "small function class" approach is more general than the subsampling approach.

Finally, one difference between our approach and the subsampling approach is the kind of results of interest. We are mainly concerned with asymptotic results, and on our way achieve approximation guarantees which are good for large sample size n. The focus of the subsampling papers is non-asymptotic, dealing with a small or moderate sample size n, and to prove approximation guarantees in this regime.

5.5. Other Statistical Learning Theory Approaches to Clustering. In the last years there have been several papers which started to look at clustering from a statistical learning theory perspective. A general statistical learning theory approach to clustering, based on a very similar intuition as ours, has already been presented in Buhmann [1998]. Here the authors put forward an "empirical risk approximation" approach for unsupervised learning, along the lines of empirical risk minimization for the supervised case. The setting under consideration is that the clustering quality function is an expectation with respect to the true underlying probability distribution, and the empirical quality function is the corresponding empirical expectation. Then, similar to the statistical learning theory for supervised learning, generalization bounds can be derived, for example using VC dimensions. Additionally, the authors discuss regularization approaches and relate them to annealing schemes for center-based clusterings.

A different approach has been investigated in Ben-David [2007]. Here the author formalizes the notion of a "cluster description scheme". Intuitively, a clustering problem can be described by a cluster description scheme of size $l \in \mathbb{N}$ if each clustering can be described using l points from the space (and perhaps some additional parameter). For instance, this is the case for center-based clusterings, where the clustering can be described by the centroids only. Ben-David then proves

generalization bounds for clustering description schemes which show that the global minimizer of the empirical quality function converges to the global minimizer of the true quality function. The proof techniques used in this chapter are very close to the ones used in standard minimum description length results.

Another class of results about K-means algorithms has been proved in Rakhlin and Caponnetto [2007]. After computing covering numbers for the underlying classes, the authors study the stability behavior of K-means. This leads to statements about the set of "almost-minimizers" (that is the set of all functions whose quality is ε close to the one of the global optimal solutions). As opposed to our results and all the other results discussed above, the main feature of this approach is that at the end of the day, one is able to make statements about the clustering functions themselves, rather than only about their quality values. In this sense, the approach in Rakhlin and Caponnetto [2007] has more powerful results, but its application is restricted to K-means type algorithms.

All approaches outlined above implicitly or explicitly rely on the same intuition as our approach: the function class needs to be "small" in order to lead to consistent clusterings. However, all previous results have some restrictions we could overcome in our approach. First of all, in the papers discussed above the quality function needs to be an expectation, and the empirical quality function is simply the empirical expectation. Here our results are more general: we neither require the quality functions to be expectations (for example, Ncut cannot be expressed as an expectation, it is a ratio of two expectations) nor do we require unbiasedness of the empirical quality function. Second, the papers discussed above make statements about global optimizers, but do not really deal with the question how such a global optimizer can be computed. The case of standard K-means shows that this is by no means simple, and in practice one has to use heuristics which discover local optima only. In contrast, we suggest a concrete algorithm (NNC) which computes the global optimum over the current function class, and hence our results not only concern abstract global minimizers which are hard to obtain, but refer to exactly the quantities which are computed by the algorithm. Finally, our algorithm has the advantage that it provides a framework for dealing with more general clustering objective functions than just center-based ones. This is not the case in the papers above.

Finally, we would like to mention that a rather general but vague discussion of some of the open issues in statistical approaches to clustering has been led in von Luxburg and Ben-David [2005]. This chapter partly solves some of the open issues raised there.

6. Discussion

This chapter is concerned with clustering algorithms which minimize certain quality functions. Our main point is that as soon as we require statistical consistency we have to work with function classes $\mathcal{F}_n$ which are "small". Our results have a similar taste as the well-known corresponding results for supervised classification. While in the domain of supervised classification practitioners are well aware of the effect of overfitting, it seems like this effect has been completely overlooked in the clustering domain.

We would like to highlight a convenient side-effect of working with small function classes. In clustering, for many objective functions the problem of finding the best partition of the discrete data set is an NP-hard problem (for example, this is the case for all balanced graph-cut objective functions). On the other side, if we restrict the function class $\mathcal{F}_n$ to have polynomial size (in n), then the trivial algorithm of evaluating all functions in $\mathcal{F}_n$ and selecting the best one is inherently polynomial. Moreover, if the small function class is "close" to the large function class, then the solution found in the small function class approximates the best solution in the unrestricted space of all clusterings.

We believe that the approach of using restricted function classes can be very promising, also from a practical point of view. It can be seen as a more controlled way of constructing approximate solutions of NP hard optimization problems than the standard approaches of local optimization or relaxation. While the effects of the latter cannot be controlled in general, we are able to control the effects of optimizing over smaller function classes by carefully selecting $\mathcal{F}_n$. This strategy circumvents the problem that solutions of local optimization or relaxation heuristics can be arbitrarily far away from the optimal solution.

The generic clustering algorithm we studied in this article is nearest neighbor clustering, which produces clusterings that are constant on small local neighborhoods. We have proved that this algorithm is statistically consistent for a large variety of popular clustering objective functions. Thus, as opposed to other clustering algorithms such as the K-means algorithm or spectral clustering, nearest neighbor clustering is guaranteed to converge to a minimizer of the true global optimum on the underlying space. This statement is much stronger than the results already known for K-means or spectral clustering. For K-means it has been proved that the global minimizer of the WSS objective function on the sample converges to a global minimizer on the underlying space (e.g., Pollard, 1981). However, as the standard K-means algorithm only discovers a local optimum on the discrete sample, this result does not apply to the algorithm used in practice. A related effect happens for spectral clustering, which is a relaxation attempting to minimize Ncut or $\mathrm{RatioCut}$. For this class of algorithms, it has been shown that under certain conditions the solution of the relaxed problem on the finite sample converges to some limit clustering. However, this limit clustering is not necessarily the optimizer of the Ncut or $\mathrm{RatioCut}$ objective function.

It is interesting to note that the problems about the existing consistency results for K-means and spectral clustering are "reverse" to each other: while for K-means we know that the global minimizer converges, but this result does not apply to the algorithm used in practice, for spectral clustering there exist consistency results for the algorithm used in practice, but these results do not relate to the global minimizer. For both cases, our consistency results represent an improvement: we have constructed an algorithm which provably converges to the true limit minimizer of WSS or Ncut, respectively. The same result also holds for a large number of alternative objective functions used for clustering.

We believe that a big advantage of our approach is that both the algorithm and the statistical analysis is not restricted to center-based algorithms only, as it has been the case for most approaches in the literature [Buhmann, 1998, Ben-David, 2007, Rakhlin and Caponnetto, 2007]. Instead, nearest neighbor clustering can be used as a baseline method to construct clusterings for any objective function. In von Luxburg et al. [2008] we have shown how nearest neighbor clustering can be implemented efficiently using branch and bound, and that in terms of quality, its results can compete with algorithms of spectral clustering (for the Ncut objective function) or K-means (for the WSS objective function). We believe that in particular for unusual objective functions for which no state of the art optimizer exists yet, nearest neighbor clustering is a promising baseline to start with. We have seen that for many commonly used objective functions, statistical guarantees for nearest neighbor clustering can be obtained, and we expect the same to be true for many more clustering objective functions.

Finally, it is a fair question how statistical consistency helps in practical applications. Is it any help in solving the big open issues in clustering, such as the question of selecting clustering algorithms for a particular data set, or selecting the number of clusters? In this generality, the answer is no. In our opinion, consistency is a *necessary* requirement which any clustering algorithm should satisfy. If an algorithm is not consistent, even with a high amount of data one cannot rely

on a clustering constructed on a finite amount of data—and this is not due to computational problems, but to inherent statistical problems. Such an algorithm cannot be trusted when constructing results on a finite sample; given another sample, it might just come up with a completely different clustering. Or, the more samples one gets, the more "trivial" the solution might become (unnormalized spectral clustering is an example for such an algorithm). In this sense, consistency is just one piece of evidence to discard unreliable clustering algorithms. In our opinion, it is very hard to come up with *sufficient* conditions about "what a good clustering algorithm is". The applications of clustering are just too diverse, and 50 years of clustering literature show that people will not agree on a unique definition of what a good clustering algorithm is. This is the reason why we believe that it is very fruitful to start by studying necessary conditions first. This chapter is meant as a contribution to this effort.

7. Proofs

In this section we concentrate all the proofs.

Proof of Theorem 8.1. The following lemma will be central in our analysis. It allows to take a supremum out of a probability.

LEMMA 8.1. *With the notation in Theorem* 8.1 *we have:*

$$\mathbb{P}(\sup_{f\in\mathcal{F}_n}|Q_n(f)-Q(f)|\geqslant\varepsilon)\leqslant 2s(\widetilde{\mathcal{F}_n},2n)\frac{\sup_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\geqslant\varepsilon/4)}{\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\leqslant\varepsilon/2)}.$$

The proof technique is similar to the one in Devroye et al. [1996], Section 12.3. The unusual term in the denominator originates in the symmetrization step. In a more standard setting where we have $\mathbb{E}Q_n = Q$, this term usually "disappears" as it can be lower bounded by $1/2$, for example using Chebyshev's inequality (e.g., Section 12.3 of Devroye et al., 1996). Unfortunately, this does not work in our more general case, as we do not assume unbiasedness and instead also allow $\mathbb{E}Q_n \neq Q$. However, note that the ratio in Lemma 8.1 essentially has the form $u_n/(1-u_n)$. Thus, as soon as the term u_n in the numerator becomes non-trivial (i.e., $u_n < 1$ or say, $u_n < 3/4$), then the denominator will only play the role of a small constant (it is lower bounded by 1/4). This means that in the regime where the numerator is non-trivial, the whole bound will essentially behave like the numerator.

PROOF. First note that we can replace the data-dependent function class $\mathcal{F}_n$ by the class $\widetilde{\mathcal{F}_n}$ which does not depend on the data:

$$\mathbb{P}(\sup_{f\in\mathcal{F}_n}|Q_n(f)-Q(f)|\geqslant\varepsilon)\leqslant\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q(f)|\geqslant\varepsilon).$$

Now we want to use a symmetrization argument. To this end, let $X'_1,\ldots,X'_n$ be a ghost sample (that is a sample drawn i.i.d. according to $\mathbb{P}$ which is independent of our first sample $X_1,\ldots,X_n$), and denote by Q'_n the empirical quality function based on the ghost sample.

Let $\widehat{f}\in\widetilde{\mathcal{F}_n}$ be such that $|Q_n(\widehat{f})-Q(\widehat{f})|\geqslant\varepsilon$; if such an $\widehat{f}$ does not exist then just choose $\widehat{f}$ as some other fixed function in $\widetilde{\mathcal{F}_n}$. Note that $\widehat{f}$ is a data-dependent function depending on the sample $X_1,...,X_n$. We have the following inequalities:

$$\begin{aligned}&\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q'_n(f)|\geqslant\varepsilon/2)\\&\geqslant\mathbb{P}(|Q_n(\widehat{f})-Q'_n(\widehat{f})|\geqslant\varepsilon/2)\end{aligned}$$

$$
\begin{aligned}
&\geqslant \mathbb{P}(|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon, |Q'_n(\widehat{f}) - Q(\widehat{f})| \leqslant \varepsilon/2) \\
&= \mathbb{E}\left(\mathbb{P}(|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon, |Q'_n(\widehat{f}) - Q(\widehat{f})| \leqslant \varepsilon/2 | X_1, \dots, X_n)\right) \\
&= \mathbb{E}\left(\mathbb{P}(|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon | X_1, \dots, X_n) \mathbb{P}(|Q'_n(\widehat{f}) - Q(\widehat{f})| \leqslant \varepsilon/2 | X_1, \dots, X_n)\right) \\
&= \mathbb{E}\left(1_{|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon} \mathbb{P}(|Q'_n(\widehat{f}) - Q(\widehat{f})| \leqslant \varepsilon/2 | X_1, \dots, X_n)\right) \\
&\geqslant \mathbb{E}\left(1_{|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon} \inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q'_n(f) - Q(f)| \leqslant \varepsilon/2 | X_1, \dots, X_n)\right) \\
&= \mathbb{E}(1_{|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon}) \mathbb{E}\left(\inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q'_n(f) - Q(f)| \leqslant \varepsilon/2 | X_1, \dots, X_n)\right) \\
&= \mathbb{P}(|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon) \inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \leqslant \varepsilon/2) \\
&= \mathbb{P}(\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q(f)| \geqslant \varepsilon) \inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \leqslant \varepsilon/2).
\end{aligned}
$$

The last step is true because of the definition of $\widehat{f}$: note that due to the definition of $\widehat{f}$ the event $|Q_n(\widehat{f}) - Q(\widehat{f})| \geqslant \varepsilon$ is true iff there exists some $f \in \widetilde{\mathcal{F}_n}$ such that $|Q_n(f) - Q(f)| \geqslant \varepsilon$, which is true iff $\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q(f)| \geqslant \varepsilon$ (recall that we assumed for ease of notations that all supremum are attained). Rearranging the inequality above leads to

$$
\mathbb{P}(\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q(f)| \geqslant \varepsilon) \leqslant \frac{\mathbb{P}(\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q'_n(f)| \geqslant \varepsilon/2)}{\inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \leqslant \varepsilon/2)}.
$$

Due to the symmetrization we got rid of the quantity $Q(f)$ in the numerator. Furthermore, using the assumption of the theorem that $Q_n(f)$ does not involve any function evaluations $f(x)$ for $x \notin \{X_1, \dots, X_n\}$ we can apply a union bound argument to move the supremum in the numerator out of the probability:

$$
\begin{aligned}
&\mathbb{P}(\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q'_n(f)| \geqslant \varepsilon/2) \\
&\leqslant s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q'_n(f)| \geqslant \varepsilon/2) \\
&\leqslant s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| + |Q(f) - Q'_n(f)| \geqslant \varepsilon/2) \\
&\leqslant 2s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \geqslant \varepsilon/4).
\end{aligned}
$$

This completes the proof of the lemma. □

We can now prove Theorem 8.1.

PROOF. Additionally to the functions f_n and f^*, we will define

$$
\begin{aligned}
f_n^* &\in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q(f), \\
\widetilde{f}^* &\in \operatorname*{argmin}_{f \in \mathcal{F}_n} d(f, f^*).
\end{aligned}
$$

To prove the theorem we have to show that under the conditions stated, for any fixed $\varepsilon > 0$ the term $\mathbb{P}(|Q(f_n) - Q(f^*)| \geqslant \varepsilon)$ converges to 0. We can study each "side" of this convergence independently:

$$\mathbb{P}(|Q(f_n) - Q(f^*)| \geqslant \varepsilon) = \mathbb{P}(Q(f_n) - Q(f^*) \leqslant -\varepsilon) + \mathbb{P}(Q(f_n) - Q(f^*) \geqslant \varepsilon).$$

To treat the "first side" observe that if $f_n \in \mathcal{F}$ then $Q(f_n) - Q(f^*) > 0$ by the definition of f^*. This leads to

$$\mathbb{P}(Q(f_n) - Q(f^*) \leqslant -\varepsilon) \leqslant \mathbb{P}(f_n \notin \mathcal{F}).$$

Under Assumption (2) of Theorem 8.1 this term tends to 0.

The main work of the proof is to take care of the second side. To this end we split $Q(f_n) - Q(f^*)$ in two terms, the estimation error and the approximation error:

$$Q(f_n) - Q(f^*) = Q(f_n) - Q(f_n^*) + Q(f_n^*) - Q(f^*).$$

For a fixed $\varepsilon > 0$ we have

$$\mathbb{P}(Q(f_n) - Q(f^*) \geqslant \varepsilon) \leqslant \mathbb{P}(Q(f_n) - Q(f_n^*) \geqslant \varepsilon/2) + \mathbb{P}(Q(f_n^*) - Q(f^*) \geqslant \varepsilon/2).$$

In the following sections we will treat both parts separately.

Estimation error. The first step is to see that

$$Q(f_n) - Q(f_n^*) \leqslant 2 \sup_{f \in \mathcal{F}_n} |Q_n(f) - Q(f)|.$$

Indeed, since $Q_n(f_n) \leqslant Q_n(f_n^*)$ by the definition of f_n we have

$$\begin{aligned} Q(f_n) - Q(f_n^*) &= Q(f_n) - Q_n(f_n) + Q_n(f_n) - Q_n(f_n^*) + Q_n(f_n^*) - Q(f_n^*) \\ &\leqslant Q(f_n) - Q_n(f_n) + Q_n(f_n^*) - Q(f_n^*) \\ &\leqslant 2 \sup_{f \in \mathcal{F}_n} |Q_n(f) - Q(f)|. \end{aligned}$$

Using Lemma 8.1 we obtain

$$\mathbb{P}(Q(f_n) - Q(f_n^*) \geqslant \varepsilon/2) \leqslant 2s(\widetilde{\mathcal{F}_n}, 2n) \frac{\sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \geqslant \varepsilon/16)}{\inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \leqslant \varepsilon/8)}.$$

Now observe that under Assumption (1) the numerator of the expression in the proposition tends to 0 and the denominator tends to 1, so the whole term tends to 0.

Approximation Error. By definition of f_n^* it is clear that

$$Q(f_n^*) - Q(f^*) \leqslant Q(\widetilde{f}^*) - Q(f^*).$$

Using Assumption (3) this leads to

$$\begin{aligned} \mathbb{P}(Q(f_n^*) - Q(f^*) \geqslant \varepsilon/2) &\leqslant \mathbb{P}(Q(\widetilde{f}^*) - Q(f^*) \geqslant \varepsilon/2) \\ &\leqslant \mathbb{P}(d(f, \widetilde{f}^*) \geqslant \delta(\varepsilon/2)). \end{aligned}$$

The right hand side clearly tends to 0 by Assumption (2). □

Proof of Theorem 8.2. Before proving Theorem 8.2, we again need to prove a few technical lemmas. The first one is a simple relation between the shattering coefficients of the nearest neighbor function classes.

LEMMA 8.2. *Let $u \in \mathbb{N}$ and $\widetilde{\mathcal{F}_n}$ and $\widehat{\mathcal{F}_n}$ be the function sets defined in Theorem* 8.2. *Then*

$$s(\widetilde{\mathcal{F}_n}, u) \leqslant s(\widehat{\mathcal{F}_n}, u) \leqslant K^m u^{(d+1)m^2}.$$

PROOF. The first inequality is obvious as we have $\widetilde{\mathcal{F}_n} \subset \widehat{\mathcal{F}_n}$. For the second inequality observe that

$$s(\widehat{\mathcal{F}_n}, u) \leqslant K^m s^*(\widehat{\mathcal{F}_n}, u)$$

where $s^*(\widehat{\mathcal{F}_n}, u)$ is the maximal number of different ways u points can be partitioned by cells of a Voronoi partition of m points. It is well known (e.g., Section 21.5 of Devroye et al., 1996) that $s^*(\widehat{\mathcal{F}_n}, u) \leqslant u^{(d+1)m^2}$ for $d > 1$. Note that for $d = 1$ a similar inequality holds, we do not consider this case any further. □

The second lemma relates a function evaluated at a point x to the same function, evaluated at the nearest neighbor of x in the training points. This lemma builds on ideas of Fritz [1975].

LEMMA 8.3. *Let $f : \mathcal{X} \to \{1, \ldots, K\}$ be continuous almost everywhere and*

$$L_n := \mathbb{P}(f(X) \neq f(NN_m(X))|X_1, \ldots, X_n).$$

Then for every $\varepsilon > 0$ there exists a constant $b_f(\varepsilon) > 0$ independent of n such that

$$\mathbb{P}(L_n \geqslant \varepsilon) \leqslant \frac{2}{\varepsilon} e^{-m b_f(\varepsilon)}.$$

PROOF. By $B(x, \delta)$ we denote the Euclidean ball of center x and radius δ. The first step of the proof consists in constructing a certain set D (depending on ε) which satisfies the following statement:

For all $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$, a measurable set $D \subset \mathbb{R}^d$ and a constant $1 > u > 0$ such that

(a) $\mathbb{P}(D) \geqslant 1 - \varepsilon/2$

(b) $\forall x \in D : \mathbb{P}(B(x, \delta)) > u$

(c) $\forall x \in D$ *the function f is constant on $B(x, \delta)$.*

Assume we have such a set D. Then using Properties (c) and (a) we can see that

$$\begin{aligned} L_n &= \mathbb{P}(f(X) \neq f(NN_m(X))|X_1, \ldots, X_n) \\ &\leqslant \mathbb{P}(X \notin D|X_1, \ldots, X_n) + \mathbb{P}(X \in D, |X - NN_m(X)| > \delta|X_1, \ldots, X_n) \\ &\leqslant \tfrac{\varepsilon}{2} + \mathbb{P}(X \in D, |X - NN_m(X)| > \delta|X_1, \ldots, X_n). \end{aligned}$$

Using the Markov inequality we can then see that

$$\begin{aligned} \mathbb{P}(L_n > \varepsilon) &\leqslant \mathbb{P}(\mathbb{P}(X \in D, |X - NN_m(X)| > \delta|X_1, \ldots, X_n) \geqslant \tfrac{\varepsilon}{2}) \\ &\leqslant \tfrac{2}{\varepsilon}\mathbb{E}(\mathbb{P}(X \in D, |X - NN_m(X)| > \delta|X_1, \ldots, X_n)) \\ &= \tfrac{2}{\varepsilon}\mathbb{P}(X \in D, |X - NN_m(X)| > \delta) \\ &= \tfrac{2}{\varepsilon}\textstyle\int_D \mathbb{P}(|x - NN_m(x)| > \delta)\, d\mathbb{P}(x). \end{aligned}$$

Due to Property (b) we know that for all $x \in D$,

$$\begin{aligned}\mathbb{P}(|x - NN_m(x)| > \delta) &= \mathbb{P}(\forall i \in \{1, \dots, m\}, x \notin B(X_i, \delta)) \\ &= (1 - \mathbb{P}(B(x, \delta)))^m \\ &\leqslant (1 - u)^m.\end{aligned}$$

Setting $b(\varepsilon) := -log(1 - u) > 0$ then leads to

$$P(L_n > \varepsilon) \leqslant \frac{2}{\varepsilon}\mathbb{P}(D)(1 - u)^m \leqslant \frac{2}{\varepsilon}e^{-mb(\delta(\varepsilon))}.$$

Note that this constant $b(\varepsilon)$ will also be used in several of the following lemmas. To finish the proof of the lemma we have to show how the set D can be constructed. By the assumption of the lemma we know that f is continuous a.e., and that f only takes finitely many values $1, ..., K$. This implies that the set

$$C = \{x \in \mathbb{R}^d : \exists \delta > 0 : d(x, y) \leqslant \delta \Rightarrow f(x) = f(y)\}$$

satisfies $\mathbb{P}(C) = 1$. Furthermore, for any $\delta > 0$ we define the set

$$A_\delta = \{x \in C : d(x, y) \leqslant \delta \Rightarrow f(x) = f(y)\}.$$

We have $\cup_\delta A_\delta = C$, and for $\sigma > \delta$ we have $A_\sigma \subset A_\delta$. This implies that given some $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$ such that $\mathbb{P}(A_{\delta(\varepsilon)}) \geqslant 1 - \varepsilon/4$. By construction, all points in $A_{\delta(\varepsilon)}$ satisfy Property (c).

As the next step, we can see that for every $\delta > 0$ one has $\mathbb{P}(B(x, \delta)) > 0$ almost surely (with respect to x). Indeed, the set $U = \{x : \exists \delta > 0 : \mathbb{P}(B(x, \delta)) = 0\}$ is a union of sets of probability zero. So using the fact that $\mathbb{R}^d$ is separable we see that $\mathbb{P}(U) = 0$. Thus, $\mathbb{P}(\mathbb{P}(B(X, \delta)|X) > 0) = 1$, which implies $\mathbb{P}(\mathbb{P}(B(X, \delta)|X) > \frac{1}{n}) \to 1$. This means that given $\varepsilon > 0$ and $\delta > 0$ there exists a set A and a constant $u > 0$ such that $\mathbb{P}(A) \geqslant 1 - \varepsilon/4$ and $\forall x \in A, \mathbb{P}(B(x, \delta)) > u$. So all points in A satisfy Property (b).

Now finally define the set $D = A \bigcap A_{\delta(\varepsilon)}$. By construction, this set has probability $P(D) \geqslant \varepsilon/2$, so it satisfies Property (a). It satisfies Properties (b) and (c) by construction of A and $A_{\delta(\varepsilon)}$, respectively. □

Proof of Theorem 8.2. To prove this theorem we will verify that the conditions (1) - (3) of Theorem 8.1 are satisfied for the function classes studied in Theorem 8.2.

Lemma 8.2 proves that Condition (1) of Theorem 8.2 implies Condition (1) of Theorem 8.1. Moreover, it is obvious that Condition (3) of Theorem 8.2 implies Condition (3) of Theorem 8.1.

Thus we only have to prove Condition (2) of Theorem 8.1. We begin by proving that $\mathbb{P}(f_n \notin \mathcal{F}) \to 0$. As $f_n \in \mathcal{F}_n$ by definition we have that $\Phi_n(f_{n,k}) > a_n$ for all $k = 1, \dots, K$. A union bound argument shows that

$$\mathbb{P}(f_n \notin \mathcal{F}) \leqslant K \sup_k \mathbb{P}(\Phi(f_{n,k}) \leqslant a).$$

Using the same techniques as in the proof of Lemma 8.1 we can see that

$$\begin{aligned}\mathbb{P}(\Phi(f_{n,k}) \leqslant a) &\leqslant \mathbb{P}(\Phi_n(f_{n,k}) - \Phi(f_{n,k}) \geqslant a_n - a) \\ &\leqslant \mathbb{P}(\sup_{g \in \mathcal{F}_n} \Phi_n(g_k) - \Phi(g_k) \geqslant a_n - a)\end{aligned}$$

$$\leqslant 2s(\widehat{\mathcal{F}_n}, 2n)\frac{\sup_{g\in\widehat{\mathcal{F}_n}} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \geqslant (a_n - a)/4)}{\inf_{g\in\widehat{\mathcal{F}_n}} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \leqslant (a_n - a)/2)}.$$

Moreover, we already proved in Lemma 8.2 that $s(\widehat{\mathcal{F}_n}, 2n) \leqslant K^m(2n)^{(d+1)m^2}$. Condition (5) of Theorem 8.2 then implies that $\mathbb{P}(\Phi(f_{n,k}) \leqslant a)$ tends to 0.

Now we have to prove that for $f \in \mathcal{F}$ the term $d(f, \mathcal{F}_n) := \min_{g\in\mathcal{F}_n} d(f, g)$ tends to 0 in probability. Let $\widetilde{f}(x) = f(NN_m(x))$. If $\widetilde{f} \in \mathcal{F}_n$ then $d(f, \mathcal{F}_n) \leqslant d(f, \widetilde{f})$, so the following holds true:

$$\mathbb{P}(d(f, \mathcal{F}_n) \geqslant \varepsilon) \leqslant \mathbb{P}(\widetilde{f} \notin \mathcal{F}_n) + \mathbb{P}(d(f, \widetilde{f}) \geqslant \varepsilon).$$

The second term on the right hand side tends to 0 because of Lemma 8.3. To deal with the first term on the right hand side, observe that

$$\mathbb{P}(\widetilde{f} \notin \mathcal{F}_n) \leqslant K \sup_k \mathbb{P}(\Phi_n(\widetilde{f}_k) \leqslant a_n).$$

Because of Condition (4), for all $\varepsilon > 0$, $f \in \mathcal{F}$ and $g \in \widehat{\mathcal{F}_n}$ there exists $\delta(\varepsilon) > 0$ such that

$$d(f, g) \leqslant \delta(\varepsilon) \Rightarrow \Phi(f_k) - \Phi(g_k) \leqslant \varepsilon.$$

Define $a_n^f := \inf_k \Phi(f_k) - a_n$. Since $f \in \mathcal{F}$ there exists N such that $n \geqslant N \Rightarrow a_n^f > 0$. For $n \geqslant N$ we have the following inequalities:

$$\begin{aligned}
&\mathbb{P}(\Phi_n(\widetilde{f}_k) \leqslant a_n) \\
&= \mathbb{P}(\Phi(f_k) - \Phi_n(\widetilde{f}_k) \geqslant \Phi(f_k) - a_n) \\
&= \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) + \Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geqslant \Phi(f_k) - a_n) \\
&\leqslant \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) \geqslant (\Phi(f_k) - a_n)/2) + \mathbb{P}(\Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geqslant (\Phi(f_k) - a_n)/2) \\
&\leqslant \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) \geqslant a_n^f/2) + \mathbb{P}(\Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geqslant a_n^f/2) \\
&\leqslant \mathbb{P}(d(f, \widetilde{f}) > \delta(a_n^f/2)) + \mathbb{P}(\sup_{g\in\widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geqslant a_n^f/2) \\
&\leqslant \tfrac{2}{\delta(a_n^f/2)} e^{-mb(\delta(a_n^f/2))} + \mathbb{P}(\sup_{g\in\widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geqslant a_n^f/2).
\end{aligned}$$

If $m \to \infty$ then the first term goes to 0. Indeed, $\delta(a_n^f/2)$ and $b(\delta(a_n^f/2))$ tend to positive constants since $f \in \mathcal{F}$ and thus $a_n^f \to \inf_k \Phi(f_k) - a > 0$. For the second term, the key step is to see that by the techniques used in the proof of Lemma 8.1 we get

$$\begin{aligned}
&\mathbb{P}(\sup_{g\in\widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geqslant a_n^f/2) \\
&\leqslant 2K^m(2n)^{(d+1)m^2} \frac{\sup_{g\in\widehat{\mathcal{F}_n}} \mathbb{P}(\Phi(g_k) - \Phi_n(g_k) \geqslant a_n^f/8)}{\inf_{g\in\widehat{\mathcal{F}_n}} \mathbb{P}(\Phi(g_k) - \Phi_n(g_k) \leqslant a_n^f/4)}.
\end{aligned}$$

Under Condition (2) this term tends to 0. □

The Proofs of the Consistency Theorems 8.3, 8.5, 8.7 and 8.8. All these theorems are applications of Theorem 8.2 to specific objective functions Q_n and Q and to specific functions Φ_n and Φ. For all of them, we individually have to check whether the conditions in Theorem 8.2 are satisfied. In this section, we do not follow the order of the Theorems. This is only due to better readability of the proofs.

In most of these proofs, we will use the McDiarmid inequality [McDiarmid, 1989], see Theorem 10.4.

Moreover, several times we will use the fact that $a_n \to a, m \to \infty$ and $\frac{m^2 \log n}{n(a-a_n)^2} \to 0$ implies that $n(a-a_n)^2 \to \infty$ and $\frac{m^2 \log n}{n} \to 0$.

Before we look at the "combined" objective functions such as Ncut, $\mathrm{RatioCut}$, WSS, we will prove some technical conditions about their "ingredients" cut, vol, $Ef_k(X)$ and WS.

LEMMA 8.4 (Conditions (2), (4), and (5) for cut, vol, $\mathbb{E}f_k(X)$, and WS). *Assume that*

$$\frac{m^2 \log n}{n(a-a_n)^2} \to 0$$

then vol, cut, $\mathbb{E}f_k(X)$ *and* WS *satisfy Conditions (2), (4) and (5) of Theorem 8.2.*

PROOF. To prove Conditions (2) and (5) we are going to use the McDiarmid inequality. Observe that if one replaces one variable X_i by a new one X_i', then vol_n changes by at most $2C/n$, cut_n changes by at most $2C/n$, $\mathrm{WS}(f_k)$ changes by at most $2C/n$, and $n_k(f)$ changes by at most $1/n$. Using the McDiarmid inequality, this implies that for all $g \in \widehat{\mathcal{F}_n}$ and $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}(|\operatorname{vol}_n(g_k) - \operatorname{vol}(g_k)| \geqslant \varepsilon) \leqslant 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|\operatorname{cut}_n(g_k) - \operatorname{cut}(g_k)| \geqslant \varepsilon) \leqslant 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|\operatorname{WS}_n(g_k) - \operatorname{WS}(g_k)| \geqslant \varepsilon) \leqslant 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|n_k(g) - \mathbb{E}g_k(X)| \geqslant \varepsilon) \leqslant 2e^{-2n\varepsilon^2}.
\end{aligned}$$

So to prove Condition (2) we have to show that

$$\forall \varepsilon > 0, K^m (2n)^{(d+1)m^2} e^{-n\varepsilon} \to 0.$$

This follows clearly from $K^m (2n)^{(d+1)m^2} e^{-n\varepsilon} = e^{-n\left(\frac{m \log K + (d+1)m^2 \log(2n)}{-n} + \varepsilon\right)}$ and $\frac{m^2 \log n}{n} \to 0$. Moreover, since $n(a-a_n)^2 \to \infty$ Condition (5) is also true.

To prove (4) for each of the objective functions, let $f, g \in \mathcal{H}$ and f_k and g_k be the corresponding cluster indicator functions for cluster k. Then we can see that

$$\begin{aligned}
|\operatorname{vol}(g_k) - \operatorname{vol}(f_k)| &= \Big| \int\int (f_k(X) - g_k(X)) s(X,Y)\, dP(X) dP(Y) \Big| \\
&\leqslant C \int_{\{f_k = g_k\}} 0\, dP(X) dP(Y) + C \int_{\{f_k = g_k\}^c} 1\, dP(X) dP(Y) \\
&= C\mathbb{P}(f_k \neq g_k) \\
&\leqslant C d(f,g),
\end{aligned}$$

$$|\operatorname{cut}(g_k) - \operatorname{cut}(f_k)| = \Big| \int\int f_k(X)(1 - f_k(Y)) s(X,Y) - g_k(X)(1 - g_k(Y)) s(X,Y)\, dP(X) dP(Y) \Big|$$

$$\begin{aligned}
&\leqslant C \int \int_{\{f=g\}^2} 0 \; dP(X)dP(Y) + C \int \int_{(\{f=g\}^2)^c} 1 \; dP(X)dP(Y) \\
&= C(1 - \mathbb{P}(f(X) = g(X))^2) \\
&= C(1 - (1 - d(f,g))^2) \\
&\leqslant 2Cd(f,g),
\end{aligned}$$

$$|\mathbb{E}f_k(X) - \mathbb{E}g_k(X)| \leqslant d(f,g),$$

$$\begin{aligned}
|\operatorname{WS}(f_k) - \operatorname{WS}(g_k)| &= | \int \int (f_k(X)f_k(Y) - g_k(X)g_k(Y))s(X,Y) \; dP(X)dP(Y) \; | \\
&\leqslant \int_{\{f=g\}^2} 0 \; dP(X)dP(Y) + C \int_{(\{f=g\}^2)^c} 1 \; dP(X)dP(Y) \\
&= C(1 - \mathbb{P}(f = g)^2) \\
&= C(1 - (1 - d(f,g))^2) \\
&\leqslant 2Cd(f,g).
\end{aligned}$$

□

Now we are going to check that the "combined" objective functions Ncut, $\mathrm{RatioCut}$, Mod, WSS, BWR satisfy the conditions of Theorem 8.2. For many of the objective functions, one important step in the proof is to separate the convergence of the whole term into the convergence of the numerator and the denominator.

LEMMA 8.5 (Condition (1) for Ncut). *Assume that*

$$\frac{m^2 \log n}{n} \to 0$$

then Ncut *satisfies Condition (*1*) of Theorem* 8.2*.*

PROOF. We first want to split the deviations of Ncut into the ones of cut and vol, respectively. To this end we want to show that for any $f \in \widetilde{\mathcal{F}_n}$

$$\{|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leqslant \tfrac{a}{2}\varepsilon\} \bigcap \{|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| \leqslant \tfrac{a}{2}\varepsilon\}$$

$$\subset \{|\tfrac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)} - \tfrac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)}| \leqslant \varepsilon\}.$$

This can be seen as follows. Assume that $|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leqslant \varepsilon$ and $|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| \leqslant \varepsilon$. If $\operatorname{vol}(f_k) \neq 0$ then we have (using the facts that $\operatorname{cut}(f_k) \leqslant \operatorname{vol}(f_k)$ and that $\operatorname{vol}_n(f_k) > a_n > a$ by definition of $\widetilde{\mathcal{F}_n}$):

$$\begin{aligned}
\tfrac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)} - \tfrac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} &= \tfrac{\operatorname{cut}_n(f_k)\operatorname{vol}(f_k) - \operatorname{cut}(f_k)\operatorname{vol}_n(f_k)}{\operatorname{vol}_n(f_k)\operatorname{vol}(f_k)} \\
&\leqslant \tfrac{(\operatorname{cut}(f_k)+\varepsilon)\operatorname{vol}(f_k) - \operatorname{cut}(f_k)(\operatorname{vol}(f_k)-\varepsilon)}{\operatorname{vol}_n(f_k)\operatorname{vol}(f_k)} \\
&= \tfrac{\varepsilon}{\operatorname{vol}_n(f_k)} \tfrac{\operatorname{cut}(f_k)+\operatorname{vol}(f_k)}{\operatorname{vol}(f_k)} \\
&\leqslant \tfrac{2\varepsilon}{a}.
\end{aligned}$$

On the other hand, if $\mathrm{vol}(f_k) = 0$ then we have $\mathrm{cut}(f_k) = 0$, which implies $\mathrm{cut}_n(f_k) \leqslant \varepsilon$ by the assumption above. Thus the following statement holds true:

$$\frac{\mathrm{cut}_n(f)}{\mathrm{vol}_n(f)} - \frac{\mathrm{cut}(f)}{\mathrm{vol}(f)} = \frac{\mathrm{cut}_n(f)}{\mathrm{vol}_n(f)} \leqslant \frac{\varepsilon}{a} \leqslant \frac{2\varepsilon}{a}.$$

Using the same technique we have the same bound for $\frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)} - \frac{\mathrm{cut}_n(f_k)}{\mathrm{vol}_n(f_k)}$, which proves our set inclusion.

Now we apply a union bound and the McDiarmid inequality. For the latter, note that if one changes one X_i then $\mathrm{cut}_n(f)$ and $\mathrm{vol}_n(f)$ will change at most by $2C/n$. Together all this leads to

$$\begin{aligned}
&\mathbb{P}(|\operatorname{Ncut}(f) - \operatorname{Ncut}_n(f)| > \varepsilon) \\
&\leqslant K \sup_k \mathbb{P}(|\frac{\mathrm{cut}_n(f_k)}{\mathrm{vol}_n(f_k)} - \frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)}| > \varepsilon/K) \\
&\leqslant K \sup_k \Big(\mathbb{P}(|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| > \frac{a}{2K}\varepsilon) + \mathbb{P}(|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| > \frac{a}{2K}\varepsilon)\Big) \\
&\leqslant 4K e^{-\frac{n a^2 \varepsilon^2}{8C^2K^2}}.
\end{aligned}$$

To finish we have to prove that

$$\forall \varepsilon > 0, K^{m+1}(2n)^{(d+1)m^2} e^{-n\varepsilon} \to 0.$$

This follows clearly from $K^{m+1}(2n)^{(d+1)m^2} e^{-n\varepsilon} = e^{-n\left(\frac{(m+1)\log K + (d+1)m^2 \log(2n)}{-n} + \varepsilon\right)}$ and $\frac{m^2 \log n}{n} \to 0$. □

LEMMA 8.6 (Condition (3) for Ncut). Ncut *satisfies Condition (*3*) of Theorem* 8.2*.*

PROOF. Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. In the proof of Lemma 8.4 we have already seen that

$$\begin{aligned}
|\operatorname{cut}(f_k) - \operatorname{cut}(g_k)| &\leqslant 2Cd(f,g), \\
|\operatorname{vol}(f_k) - \operatorname{vol}(g_k)| &\leqslant 2Cd(f,g).
\end{aligned}$$

If $\mathrm{vol}(g) \neq 0$ then we have (using the fact that we always have $\mathrm{cut}(f) \leqslant \mathrm{vol}(f)$):

$$\begin{aligned}
\frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)} - \frac{\mathrm{cut}(g_k)}{\mathrm{vol}(g_k)} &= \frac{\mathrm{cut}(f_k)\,\mathrm{vol}(g_k) - \mathrm{cut}(g_k)\,\mathrm{vol}(f_k)}{\mathrm{vol}(f_k)\,\mathrm{vol}(g_k)} \\
&\leqslant \frac{(\mathrm{cut}(g_k) + 2Cd(f,g))\,\mathrm{vol}(\widetilde{f}) - \mathrm{cut}(g_k)(\mathrm{vol}(g_k) - 2Cd(f,g))}{\mathrm{vol}(f_k)\,\mathrm{vol}(g_k)} \\
&= \frac{2Cd(f,g)}{\mathrm{vol}(f_k)} \frac{\mathrm{vol}(g_k) + \mathrm{cut}(g_k)}{\mathrm{vol}(g_k)} \\
&\leqslant \frac{4C}{a} d(f,g)).
\end{aligned}$$

On the other hand if $\mathrm{vol}(g_k) = 0$ then we have $|\operatorname{cut}(f_k)| \leqslant |\operatorname{vol}(f_k)| \leqslant 2Cd(f,g)$, in which case the following holds true:

$$\frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)} - \frac{\mathrm{cut}(g_k)}{\mathrm{vol}(\widetilde{f_k})} = \frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)} \leqslant \frac{2Cd(f,g)}{a} \leqslant \frac{4C}{a} d(f,g).$$

So all in all we have

$$\operatorname{Ncut}(f) - \operatorname{Ncut}(g) \leqslant \frac{4CK}{a} d(f,g).$$

We can use the same technique to bound $\mathrm{Ncut}(g) - \mathrm{Ncut}(f)$. This proves that Ncut is Lipschitz and thus uniformly continuous. □

LEMMA 8.7 (Condition (1) for RatioCut). *Assume that*

$$\frac{m^2 \log n}{n} \to 0$$

then RatioCut *satisfies Condition* (1) *of Theorem* 8.2.

PROOF. Using exactly the same proof as for Lemma 8.5 (just changing $\mathrm{vol}_n(f_k)$ to n_k and $\mathrm{vol}(f_k)$ to $\mathbb{E}f_k(X)$ and using the fact that $\mathrm{cut}(f_k) \leqslant C\mathbb{E}f_k(X)$) we get

$$\mathbb{P}(|\operatorname{RatioCut}_n(f) - \operatorname{RatioCut}(f)| > \varepsilon)$$

$$\leqslant K \sup_k \left(\mathbb{P}(|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| > \frac{a}{(S+1)K}\varepsilon) + \mathbb{P}(|n_k(f) - \mathbb{E}f_k(X)| > \frac{a}{(S+1)K}\varepsilon) \right).$$

Now a simple McDiarmid argument (using again the fact that changing one X_i changes cut_n by at most $2S/n$) gives

$$\mathbb{P}(|\operatorname{RatioCut}_n(f) - \operatorname{RatioCut}(f)| > \varepsilon) \leqslant 2Ke^{-\frac{na^2\varepsilon}{8C^2K^2}} + 2Ke^{-\frac{na^2\varepsilon^2}{2K^2}}.$$

We conclude the proof with the same argument as in Lemma 8.5. □

LEMMA 8.8 (Condition (3) for RatioCut). RatioCut *satisfies Condition (3) of Theorem* 8.2.

PROOF. This follows by the same proof as Lemma 8.5, just changing $\mathrm{vol}_n(f_k)$ to n_k, $\mathrm{vol}(f_k)$ to $\mathbb{E}f_k(X)$ and using the fact that $\mathrm{cut}(f_k) \leqslant C\mathbb{E}f_k(X)$. □

LEMMA 8.9 (Condition (1) for BWR). *If* $m^2 \log n/n \to 0$, *then* BWR *satisfies Condition (1) of Theorem* 8.2.

PROOF. Let $f \in \widetilde{\mathcal{F}_n}$. Let $\varepsilon \leqslant a/2$. If $|\operatorname{WS}_n(f_k) - \operatorname{WS}(f_k)| \leqslant \varepsilon$ and $|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leqslant \varepsilon$ then $\operatorname{WS}(f_k) \geqslant a/2 > 0$ (because $\operatorname{WS}_n(f_k) > a_n > a$ since $f \in \widetilde{\mathcal{F}_n}$). This implies

$$\begin{aligned}
\tfrac{\operatorname{cut}_n(f_k)}{\operatorname{WS}_n(f_k)} - \tfrac{\operatorname{cut}(f_k)}{\operatorname{WS}(f_k)} &= \tfrac{\operatorname{WS}(f_k)\operatorname{cut}_n(f_k) - \operatorname{WS}_n(f_k)\operatorname{cut}(f_k)}{\operatorname{WS}_n(f_k)\operatorname{WS}(f_k)} \\
&\leqslant \tfrac{\operatorname{WS}(f_k)(\operatorname{cut}(f_k)+\varepsilon) - (\operatorname{WS}(f_k)-\varepsilon)\operatorname{cut}(f_k)}{\operatorname{WS}_n(f_k)\operatorname{WS}(f_k)} \\
&= \tfrac{\varepsilon}{\operatorname{WS}_n(f_k)} \tfrac{\operatorname{WS}(f_k)+\operatorname{cut}(f_k)}{\operatorname{WS}(f_k)} \\
&\leqslant \tfrac{2C\varepsilon}{a^2}.
\end{aligned}$$

The analogous statement holds for $\frac{\operatorname{cut}(f_k)}{\operatorname{WS}(f_k)} - \frac{\operatorname{cut}_n(f_k)}{\operatorname{WS}_n(f_k)}$. Thus, if $\varepsilon \leqslant C/a$ then

$$\begin{aligned}
&\{|\operatorname{WS}_n(f_k) - \operatorname{WS}(f_k)| \leqslant a^2\varepsilon/(2C)\} \cap \{|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leqslant a^2\varepsilon/(2C)\} \\
&\subset \{\left|\frac{\operatorname{cut}_n(f_k)}{\operatorname{WS}_n(f_k)} - \frac{\operatorname{cut}(f_k)}{\operatorname{WS}(f_k)}\right| \leqslant \varepsilon\}.
\end{aligned}$$

As a consequence, if $\varepsilon \leqslant CK/a$ we have

$$\begin{aligned}
&\mathbb{P}(|\operatorname{BWR}_n(f) - \operatorname{BWR}(f)| > \varepsilon) \leqslant K \sup_k \mathbb{P}\left(\left|\frac{\operatorname{cut}_n(f_k)}{\operatorname{WS}_n(f_k)} - \frac{\operatorname{cut}(f_k)}{\operatorname{WS}(f_k)}\right| > \varepsilon/K\right) \\
&\leqslant K \sup_k \left(\mathbb{P}(|\operatorname{WS}_n(f_k) - \operatorname{WS}(f_k)| > a^2\varepsilon/(2CK)) + \mathbb{P}(|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| > a^2\varepsilon/(2CK))\right).
\end{aligned}$$

Using the McDiarmid inequality together with the fact that changing one point changes cut_n and WS_n by at most $C/(2n)$, we get for $\varepsilon \leqslant CK/a$:

$$\mathbb{P}(|\operatorname{BWR}_n(f) - \operatorname{BWR}(f)| > \varepsilon) \leqslant 4Ke^{-\frac{na^4\varepsilon^2}{8C^4K^2}}.$$

On the other hand, for $\varepsilon > CK/a$ we have

$$\begin{aligned}
&\mathbb{P}(|\operatorname{BWR}_n(f) - \operatorname{BWR}(f)| > \varepsilon) \\
&\leqslant \mathbb{P}(|\operatorname{BWR}_n(f) - \operatorname{BWR}(f)| > SK/a) \\
&\leqslant 4Ke^{-\frac{na^4(SK/a)^2}{8C^4K^2}}.
\end{aligned}$$

So all in all we have proved that

$$\mathbb{P}(|\operatorname{BWR}_n(f) - \operatorname{BWR}(f)| > \varepsilon) \leqslant 2Ke^{-\frac{na^4(\min(\varepsilon, CK/a))^2}{8C^4K^2}}.$$

We conclude the proof with the same argument as in Lemma 8.5. □

LEMMA 8.10 (Condition (3) for BWR). BWR *satisfies Condition (*3*) of Theorem* 8.2.

PROOF. Let $\varepsilon > 0$, $f \in \mathcal{F}$ and $g \in \widetilde{\mathcal{F}_n}$. We have already proved the two following inequalities (in the proofs of Lemmas 8.4 and 8.6):

$$|\operatorname{cut}(f_k) - \operatorname{cut}(g_k)| \leqslant 2Cd(f,g),$$

$$|\operatorname{WS}(f_k) - \operatorname{WS}(g_k)| \leqslant 2Cd(f,g).$$

If $2Cd(f,g) \leqslant a/2$, then using that $\operatorname{WS}(f_k) > a$ we get $\operatorname{WS}(g_k) \geqslant a/2 > 0$. By the same technique as at the beginning of Lemma 8.9 we get

$$|\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leqslant \frac{2CK}{a^2} 2Cd(f,g).$$

Written a bit differently,

$$d(f,g) \leqslant \frac{a}{4C} \Rightarrow |\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leqslant \frac{4C^2K}{a^2} d(f,g).$$

Now recall that we want to prove that there exists $\delta > 0$ such that $d(f,g) \leqslant \delta \Rightarrow |\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leqslant \varepsilon$.

If $\varepsilon \leqslant CK/a$ then we have:

$$d(f,g) \leqslant \frac{a^2}{4C^2K}\varepsilon \leqslant \frac{a}{4C} \Rightarrow |\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leqslant \frac{4C^2K}{a^2} d(f,g) \leqslant \varepsilon.$$

On the other hand, if $\varepsilon > CK/a$ then

$$d(f,g) \leqslant \frac{a}{4C} \Rightarrow |\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leqslant \frac{4C^2K}{a^2} d(f,g) \leqslant CK/a \leqslant \varepsilon$$

so we have proved the lemma. □

LEMMA 8.11 (Condition (1) for WSS). *If* $\frac{m^2 \log n}{n} \to 0$ *and that* $\operatorname{supp} \mathbb{P} \subset B(0,A)$*, then* WSS *satisfies Condition (*1*) of Theorem* 8.2.

PROOF. Let $f \in \widetilde{\mathcal{F}_n}$. First note that

$$|\operatorname{WSS}_n(f) - \operatorname{WSS}(f)| = \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_{k,n}\|^2 - \mathbb{E} \sum_{k=1}^{K} f_k(X) \|X - c_k\|^2 \right|$$

$$\leqslant \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_{k,n}\|^2 - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_k\|^2 \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_k\|^2 - \mathbb{E} \sum_{k=1}^{K} f_k(X) \|X - c_k\|^2 \right|.$$

Now we will bound the probability for each of the terms on the right hand side. For the second term we can simply apply McDiarmid's inequality. Due to the assumption that $\operatorname{supp} \mathbb{P} \subset B(0, A)$ we know that for any two points $x, y \in \operatorname{supp} \mathbb{P}$ we have $\|x - y\| \leqslant 2A$. Thus if one changes one variable X_i then the term $\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_k\|^2$ will change by at most $A^2/(4n)$. This leads to

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_k\|^2 - \mathbb{E} \sum_{k=1}^{K} f_k(X) \|X - c_k\|^2 \right| \geqslant \varepsilon \right) \leqslant 2e^{-\frac{2n\varepsilon^2}{A^4}}.$$

Now we have to take care of the first term, which can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \left(\|X_i - c_{k,n}\|^2 - \|X_i - c_k\|^2 \right).$$

The triangle inequality gives

$$\|X_i - c_{k,n}\|^2 \leqslant (\|X_i - c_k\| + \|c_{k,n} - c_k\|)^2,$$

and together with the fact that $\operatorname{supp} \mathbb{P} \subset B(0, A)$ this leads to

$$\|X_i - c_{k,n}\|^2 - \|X_i - c_k\|^2 \leqslant 6A \|c_{k,n} - c_k\|.$$

So at this point we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_{k,n}\|^2 - \|X_i - c_k\|^2 \right| \leqslant 6A \sup_k \|c_{k,n} - c_k\|.$$

We will denote the j-th coordinate of a vector X by X^j. Recall that d denotes the dimensionality of our space. Using this notation we have

$$\|c_{k,n} - c_k\|^2 = \sum_{j=1}^{d} \left(\frac{\mathbb{E} f_k(X) X^j}{\mathbb{E} f_k(X)} - \frac{1}{n_k} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i^j \right)^2.$$

Our goal will be to apply the McDiarmid inequality to each coordinate. Before we can do this, we want to show that

$$\{|n_k - \mathbb{E} f_k(X)| \leqslant \tfrac{a\varepsilon}{A+1}\} \cap \{|\tfrac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i^j - \mathbb{E} f_k(X) X^j| \leqslant \frac{a\varepsilon}{A+1}\} \subset \{|c_k^j - c_{k,n}^j| \leqslant \varepsilon\}.$$

To this end, assume that $|n_k - \mathbb{E} f_k(X)| \leqslant \varepsilon$ and $|\frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i^j - \mathbb{E} f_k(X) X^j| \leqslant \varepsilon$.

In case $\mathbb{E} f_k(X) \neq 0$ we have

$$\begin{aligned} c_k^j - c_{k,n}^j &= \frac{n_k \mathbb{E} f_k(X) X^j - \mathbb{E} f_k(X) \frac{1}{n_k} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i^j}{n_k \mathbb{E} f_k(X)} \\ &\leqslant \frac{(\mathbb{E} f_k(X) + \varepsilon) \mathbb{E} f_k(X) X^j - \mathbb{E} f_k(X) (\mathbb{E} f_k(X) X^j - \varepsilon)}{n_k \mathbb{E} f_k(X)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\varepsilon}{n_k} \frac{\mathbb{E} f_k(X) X^j + \mathbb{E} f_k(X)}{\mathbb{E} f_k(X)} \\
&\leqslant \frac{(A+1)\varepsilon}{a}
\end{aligned}
$$

and similarly for $c_{k,n}^j - c_k^j$.

On the other hand, in case $\mathbb{E} f_k(X) = 0$ we also have $\mathbb{E} f_k(X) X^j = 0$ (as f_k is a non-negative function and $|X|$ is bounded by A). Together with the assumption this means that $\frac{1}{n}\sum_{i=1}^n f_k(X_i) X_i^j \leqslant \varepsilon$. This implies

$$
|c_k^j - c_{k,n}^j| = \frac{1}{n_k} \frac{1}{n} \sum_{i=1}^n f_k(X_i) X_i^j \leqslant \frac{\varepsilon}{a} \leqslant \frac{(A+1)\varepsilon}{a}
$$

which shows the inclusion stated above. The McDiarmid inequality now yields the two statements

$$
\begin{aligned}
&\mathbb{P}(|n_k - \mathbb{E} f_k(X)| > \varepsilon) \leqslant 2e^{-2n\varepsilon^2}, \\
&\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n f_k(X_i) X_i^j - \mathbb{E} f_k(X) X^j \right| > \varepsilon \right) \leqslant 2e^{-\frac{2n\varepsilon^2}{A^2}}.
\end{aligned}
$$

Together they show that for the coordinate-wise differences

$$
\mathbb{P}(|c_k^j - c_{k,n}^j| > \varepsilon) \leqslant 2e^{-\frac{2na^2\varepsilon^2}{(A+1)^2}} + 2e^{-\frac{2na^2\varepsilon^2}{A^2(A+1)^2}} \leqslant 4e^{-\frac{2na^2\varepsilon^2}{\max(1,A^2)(A+1)^2}}.
$$

This leads to

$$
\begin{aligned}
\mathbb{P}(\|c_k - c_{k,n}\| > \varepsilon) = \mathbb{P}(\sum_{j=1}^d |c_k^j - c_{k,n}^j|^2 > \varepsilon^2) &\leqslant d \sup_j \mathbb{P}(|c_k^j - c_{k,n}^j| > \varepsilon/\sqrt{d}) \\
&\leqslant 4de^{-\frac{2na^2\varepsilon^2}{d\max(1,A^2)(A+1)^2}}.
\end{aligned}
$$

Combining all this leads to a bound for the first term of the beginning of the proof:

$$
\begin{aligned}
&\mathbb{P}\left(\left| \tfrac{1}{n}\sum_{i=1}^n \sum_{k=1}^K f_k(X_i) \big(\|X_i - c_{k,n}\|^2 - \|X_i - c_k\|^2 \big) \right| \geqslant \varepsilon \right) \\
&\leqslant \mathbb{P}\left(\sup_k \|c_{k,n} - c_k\| \geqslant \varepsilon/(6A) \right) \\
&\leqslant K \sup_k \mathbb{P}(\|c_{k,n} - c_k\| \geqslant \varepsilon/(6A)) \\
&\leqslant 4dKe^{-\frac{na^2\varepsilon^2}{18d\max(1,A^2)A^2(A+1)^2}}.
\end{aligned}
$$

Now we combine the probabilities for the first and the second term from the beginning of the proof using a union bound to get

$$
\mathbb{P}(|\operatorname{WSS}_n(f) - \operatorname{WSS}(f)| > \varepsilon) \leqslant 4dKe^{-\frac{na^2\varepsilon}{18d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{8n\varepsilon^2}{A^4}}.
$$

We conclude the proof with the same argument as in Lemma 8.5. □

LEMMA 8.12 (Condition (3) for WSS). *Assume that* $\operatorname{supp} \mathbb{P} \subset B(0, A)$ *then* WSS *satisfies Condition (3) of Theorem* 8.2.

PROOF. Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. We begin with the following inequality, which can be seen by splitting the expectation in the part where $\{f = g\}$ and $\{f \neq g\}$ and using the fact that

$\operatorname{supp}\mathbb{P} \subset B(0,A)$:

$$\begin{aligned} |\operatorname{WSS}(f) - \operatorname{WSS}(g)| &= |\mathbb{E}\textstyle\sum_{k=1}^{K} f_k(X)\|X - c_k(f)\|^2 - g_k(X)\|X - c_k(g)\|^2| \\ &\leqslant 4A^2 d(f,g) + \textstyle\int_{\{f=g\}} \sum_{k=1}^{K} f_k(X)\left(\|X - c_k(f)\|^2 - \|X - c_k(g)\|^2\right). \end{aligned}$$

For the second term we have already seen in the proof of the previous lemma that $\|X - c_k(f)\|^2 - \|X - c_k(g)\|^2 \leqslant 6A\|c_k(f) - c_k(g)\|$. So for the moment we have

$$|\operatorname{WSS}(f) - \operatorname{WSS}(g)| \leqslant 4A^2 d(f,g) + 6A \sup_k \|c_k(f) - c_k(g)\|.$$

Now we want to bound the expression $\|c_k(f) - c_k(g)\|$. First of all, observe that $|\mathbb{E}f_k(X) - g_k(X)| \leqslant d(f,g)$ and $\|\mathbb{E}f_k(X)X - g_k(X)X\| \leqslant Ad(f,g)$.

In case $\mathbb{E}g_k(X) \neq 0$ we have

$$\begin{aligned} \|c_k(f) - c_k(g)\| &= \frac{\|\mathbb{E}g_k(X)\mathbb{E}f_k(X)X - \mathbb{E}f_k(X)\mathbb{E}g_k(X)X\|}{\mathbb{E}f_k(X)\mathbb{E}g_k(X)} \\ &\leqslant \frac{\|\mathbb{E}g_k(X)\left(\mathbb{E}f_k(X)X - \mathbb{E}g_k(X)X\right)\| + \|\left(\mathbb{E}g_k(X) - \mathbb{E}f_k(X)\right)\mathbb{E}g_k(X)X\|}{\mathbb{E}f_k(X)\mathbb{E}g_k(X)} \\ &\leqslant \frac{\mathbb{E}g_k(X)\|\mathbb{E}f_k(X)X - g_k(X)X\| + A\mathbb{E}g_k(X)|\mathbb{E}g_k(X) - f_k(X)|}{\mathbb{E}f_k(X)\mathbb{E}g_k(X)} \\ &\leqslant \frac{2A}{\mathbb{E}f_k(x)} d(f,g) \\ &\leqslant \frac{2A}{a} d(f,g). \end{aligned}$$

On the other hand, in case $\mathbb{E}g_k(X) = 0$ we also have $\mathbb{E}g_k(X)X = 0$ (as g_k is a non-negative function and $|X|$ is bounded by A). This leads to

$$\|c_k(f) - c_k(g)\| = \|\frac{\mathbb{E}f_k(X)X}{\mathbb{E}f_k(X)} - \frac{\mathbb{E}g_k(X)X}{\mathbb{E}g_k(X)}\| = \|\frac{\mathbb{E}f_k(X)X}{\mathbb{E}f_k(X)}\| \leqslant \frac{A}{a} d(f,g) \leqslant \frac{2A}{a} d(f,g).$$

Combining all results leads to

$$|\operatorname{WSS}(f) - \operatorname{WSS}(g)| \leqslant 4A^2(1 + 3/a)d(f,g)$$

which proves the lemma. □

LEMMA 8.13 (Condition (1) for Mod). *If* $m^2 \log n/n \to 0$, *then* Mod *satisfies Condition (1) of Theorem* 8.2.

PROOF. Let $f \in \widetilde{f}$. Using McDiarmid inequality one can prove

$$\mathbb{P}(|\sum_{k=1}^{K} \frac{1}{n(n-1)} \sum_{i \neq j} f_k(X_i) f_k(X_j) s(X_i, X_j) - \sum_{k=1}^{K} \mathbb{E}f_k(X) f_k(Y) s(X,Y)| \geqslant \varepsilon) \leqslant 2e^{-\frac{n\varepsilon^2}{2C^2K^2}}.$$

Now for ease of notation let

$$Q_n(f) = \frac{1}{n(n-1)^3} \sum_{k=1}^{K} \sum_{i \neq j} f_k(X_i) f_k(X_j) \sum_{l, l \neq i} s(X_i, X_l) \sum_{l, l \neq j} s(X_j, X_l),$$

$$\widetilde{Q_n}(f) = \frac{1}{n(n-1)} \sum_{k=1}^{K} \sum_{i \neq j} f_k(X_i) f_k(X_j) \int s(X_i, Z) d\mathbb{P}(Z) \int s(X_j, Z) d\mathbb{P}(Z),$$

$$Q(f) = \sum_{k=1}^{K} \int \int f_k(X) f_k(Y) \int s(X,Z) d\mathbb{P}(Z) \int s(Y,Z) d\mathbb{P}(Z) d\mathcal{P}(X,Y).$$

If we have an exponential bound for $\mathbb{P}(|Q_n(f) - Q(f)| \geqslant \varepsilon)$ then with the above bound we would have an exponential bound for $\mathbb{P}(|\operatorname{Mod}_n(f) - \operatorname{Mod}(f)| \geqslant \varepsilon)$. Thus with the same argument than the one at the end of Lemma 8.5 the current lemma will be proved.
First note that

$$\mathbb{P}(|Q_n(f) - Q(f)| \geqslant \varepsilon) \leqslant \mathbb{P}(|Q_n(f) - \widetilde{Q_n}(f)| \geqslant \varepsilon/2) + \mathbb{P}(|\widetilde{Q_n}(f) - Q(f)| \geqslant \varepsilon/2).$$

Moreover $\mathbb{E}\widetilde{Q_n}(f) = Q(f)$ and thus with McDiarmid one can prove that

$$\mathbb{P}(|\widetilde{Q_n}(f) - Q(f)| \geqslant \varepsilon) \leqslant 2e^{\frac{n\varepsilon^2}{2C^4K^2}}.$$

The next step is to use the fact that for real numbers $a, b, a_n, b_n \in B(0, C)$,

$$|ab - a_nb_n| = |ab - a_nb + a_nb - a_nb_n| \leqslant C(|a - a_n| + |b - b_n|).$$

This implies the following inequalities:

$$\begin{aligned}
&|Q_n(f) - \widetilde{Q_n}(f)| \\
&\leqslant \frac{K}{n(n-1)} \sum_{i \neq j} \left| \frac{1}{(n-1)^2} \sum_{l,l\neq i} s(X_i, X_l) \sum_{l,l\neq j} s(X_j, X_l) - \int s(X_i, Z)d\mathbb{P}(Z) \int s(X_j, Z)d\mathbb{P}(Z) \right| \\
&\leqslant 2CK \sup_i \left| \frac{1}{n-1} \sum_{l,l\neq i} s(X_i, X_l) - \int s(X_i, Z)d\mathbb{P}(Z) \right|.
\end{aligned}$$

Hence the following:

$$\begin{aligned}
\mathbb{P}(|Q_n(f) - \widetilde{Q_n}(f)| \geqslant \varepsilon) &\leqslant \mathbb{P}(\sup_i |\frac{1}{n-1} \sum_{l,l\neq i} s(X_i, X_l) - \int s(X_i, Z)d\mathbb{P}(Z)| \geqslant \varepsilon/(2CK)) \\
&\leqslant n \sup_i \mathbb{P}(|\frac{1}{n-1} \sum_{l,l\neq i} s(X_i, X_l) - \int s(X_i, Z)d\mathbb{P}(Z)| \geqslant \varepsilon/(2CK)).
\end{aligned}$$

Now to bound the last term we condition on X_i and use the McDiarmid inequality. Then taking the expectation yields the exponential bound:

$$\begin{aligned}
&\mathbb{P}(|\frac{1}{n-1} \sum_{l,l\neq i} s(X_i, X_l) - \int s(X_i, Z)d\mathbb{P}(Z)| \geqslant \varepsilon/(2CK)) \\
&= \mathbb{E}(\mathbb{P}(|\frac{1}{n-1} \sum_{l,l\neq i} s(X_i, X_l) - \int s(X_i, Z)d\mathbb{P}(Z)| \geqslant \varepsilon/(2CK)|X_i)) \\
&\leqslant \mathbb{E}(2e^{-\frac{n\varepsilon^2}{2C^4K^2}}) \\
&= 2e^{-\frac{n\varepsilon^2}{2C^4K^2}}.
\end{aligned}$$

All in all we proved that

$$\mathbb{P}(|\operatorname{Mod}_n(f) - \operatorname{Mod}(f)| \geqslant \varepsilon) \leqslant 2e^{-\frac{n\varepsilon^2}{8C^2K^2}} + 2(n+1)e^{-\frac{n\varepsilon^2}{32C^2K^2}}.$$

The n in front of the exponential obviously does not matter for the limit, see end of the proof of Lemma 8.5. □

LEMMA 8.14 (Condition (3) for Mod). Mod *satisfies Condition (3) of Theorem* 8.2.

PROOF. Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. Following the proof of Lemma 8.6 we have:

$$\begin{aligned} |\operatorname{Mod}(f) - \operatorname{Mod}(g)| &\leqslant \sum_{k=1}^{K} \int \int_{(\{f=g\}^2)^c} (C + C^2) \\ &= K(C + C^2)(1 - (1 - d(f,g))^2) \\ &\leqslant 2K(C + C^2) d(f,g). \end{aligned}$$

□

The Proofs of the Convergence Rates in Theorems 8.4 **and** 8.6 The following lemma collects all the bounds given in the previous proofs for WSS . Whenever possible, we used the one-sided McDiarmid inequality.

LEMMA 8.15. *Assume that* $\operatorname{supp} \mathbb{P} \subset B(0, A)$ *for some constant* $A > 0$. *Let* $a_n^* := \inf_k \mathbb{E} f_k^*(X) - a_n$. *Then* $a_n^* \to a^* := \inf_k \mathbb{E} f_k^*(X) - a > 0$. *For all* n *and* $\varepsilon > 0$ *there exists a constant* $b(a_n^*/2)$ *which tends to a constant* $C' > 0$ *when* $n \to \infty$, *and a constant* $b(\varepsilon/(8A^2(1+3/a)))$ *(see Lemma* 8.3 *for more details about* b*) such that the following holds true*

$$\mathbb{P}(|\operatorname{WSS}(f_n) - \operatorname{WSS}(f^*)| \geqslant \varepsilon)$$

$$\leqslant 2K^{m+1}(2n)^{(d+1)m^2} \left(\frac{4dKe^{-\frac{na^2\varepsilon}{616d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{n\varepsilon^2}{32A^4}}}{1 - 4dKe^{-\frac{na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{n\varepsilon^2}{8A^4}}} + \frac{Ke^{-\frac{n(a_n-a)^2}{8}}}{1 - e^{-\frac{n(a_n-a)^2}{2}}} + \frac{Ke^{-\frac{na_n^{*2}}{32}}}{1 - e^{-\frac{na_n^{*2}}{8}}} \right)$$

$$+ \frac{4K}{a_n^*} e^{-mb(a_n^*/2)} + (16A^2(1 + 3/a)/\varepsilon) e^{-mb(\varepsilon/(8A^2(1+3/a)))}.$$

Proof of Theorem 8.4. First we take care of the last two terms. There exists N' which depends on the rate of convergence of a_n and on a^* such that for $n \geqslant N'$ we have

$$a_n^* \leqslant a^*/2.$$

This implies $b(a_n^*/2) \leqslant b(a^*/4)$ (see Lemma 8.3 for details). Now let $C_1' := b(\varepsilon/(8A^2(1+3/a)))$ and $C_2' := b(a^*/4)$. Then for $n \geqslant N'$ we have:

$$\begin{aligned} &\frac{4K}{a_n^*} e^{-mb(a_n^*/2)} + (16A^2(1 + 3/a)/\varepsilon) e^{-mb(\varepsilon/(8A^2(1+3/a)))} \\ &\leqslant 8Ka^* e^{-C_2' m} + (16A^2(1 + 3/a)/\varepsilon) e^{-C_1' m} \\ &\leqslant C_1 e^{-C_2 m} \end{aligned}$$

with

$$C_1 := \max(8Ka^*; 16A^2(1 + 3/a)/\varepsilon) \qquad \text{and} \qquad C_2 := \min(C_1'; C_2').$$

C_2 is a positive constant which depends on a, a^*, A, ε and $\mathbb{P}$. C_1 depends on K, a, a^*, ε and A.

Since we assume $n(a_n - a)^2 \to \infty$ there exists N'' which depends on the rate of convergence of a_n and on a^* such that $n \geqslant N''$ implies:

$$e^{-\frac{n(a_n-a)^2}{8}} \leqslant 1/2 \qquad \text{and} \qquad e^{-\frac{na_n^*}{32}} \leqslant e^{-\frac{n(a_n-a)^2}{8}}.$$

This means that for $n \geqslant N''$:

$$\frac{Ke^{-\frac{n(a_n-a)^2}{8}}}{1 - e^{-\frac{n(a_n-a)^2}{2}}} + \frac{Ke^{-\frac{na_n^{*2}}{32}}}{1 - e^{-\frac{na_n^{*2}}{8}}} \leqslant 4Ke^{-\frac{n(a_n-a)^2}{8}}.$$

Finally let $N = \max(N', N'')$ and

$$C_3 := \frac{8dK}{1 - 4dKe^{-\frac{Na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{N\varepsilon^2}{8A^4}}}$$
$$C_4 := \min(\frac{a^2}{616d\max(1,A^2)A^2(A+1)^2}; \frac{1}{32A^4}).$$

Since $\varepsilon \leqslant 1$ we have with these notations for $n \geqslant N$:

$$\frac{4dKe^{-\frac{na^2\varepsilon}{616d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{n\varepsilon^2}{32A^4}}}{1 - 4dKe^{-\frac{na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{n\varepsilon^2}{8A^4}}} \leqslant (C_3/2)e^{-C_4\varepsilon^2 n}.$$

All in all Theorem 8.4 is proved. □

The **Proof of Theorem 8.6** works analogously, we just replace the above lemma by to following one:

LEMMA 8.16. *Assume that the similarity function s is bounded by $C > 0$. Let $a_n^* := \inf_k \mathrm{vol}(f_k^*) - a_n$. Then $a_n^* \to \inf_k \mathrm{vol}(f_k^*) - a > 0$. For all n and $\varepsilon > 0$ there exists a constant $b(a_n^*/(2S))$ which tends to a constant $C' > 0$ when $n \to \infty$, and a constant $b(a\varepsilon/(8SK))$ (see Lemma 8.3 for more details about b) such that the following holds true*

$$\mathbb{P}(|\operatorname{Ncut}(f_n) - \operatorname{Ncut}(f^*)| \geqslant \varepsilon)$$
$$\leqslant 2K^{m+1}(2n)^{(d+1)m^2}\left(\frac{4e^{-\frac{na^2\varepsilon^2}{2048C^2K^2}}}{1-4Ke^{-\frac{na^2\varepsilon^2}{512C^2K^2}}} + \frac{e^{-\frac{n(a_n-a)^2}{32C^2}}}{1-e^{-\frac{n(a_n-a)^2}{8C^2}}} + \frac{e^{-\frac{na_n^{*2}}{128C^2}}}{1-e^{-\frac{na_n^{*2}}{32C^2}}}\right)$$
$$+\frac{4CK}{a_n^*}e^{-mb(a_n^*/(2C))} + \frac{16CK}{a\varepsilon}e^{-mb(a\varepsilon/(8CK))}.$$

CHAPTER 9

How the initialization affects the stability of the k-means algorithm

We investigate the role of the initialization for the stability of the k-means clustering algorithm. As opposed to other papers, we consider the actual k-means algorithm and do not ignore its property of getting stuck in local optima. We are interested in the actual clustering, not only in the costs of the solution. We analyze when different initializations lead to the same local optimum, and when they lead to different local optima. This enables us to prove that it is reasonable to select the number of clusters based on stability scores.

Contents

This chapter is a joint work with Ulrike Von Luxburg and Marina Meila.

1. Introduction

Stability is a popular tool for model selection in clustering, in particular to select the number k of clusters. The general idea is that the best parameter k for a given data set is the one which leads to the "most stable" clustering results. While model selection based on clustering stability is widely used in practice, its behavior is still not well-understood from a theoretical point of view. A recent line of papers discusses clustering stability with respect to the k-means criterion in an idealized setting [Ben-David et al., 2006, 2007, Shamir and Tishby, 2008a, Ben-David and von Luxburg, 2008, Shamir and Tishby, 2008b,c]. It is assumed that one has access to an ideal algorithm which can globally optimize the k-means criterion. For this perfect algorithm, results on stability are proved in the limit of the sample size n tending to infinity. However, none of these results applies to the k-means algorithm as used in practice: they do not take into account the problem of getting stuck in local optima. In this chapter we try to overcome this shortcoming. We study the stability of the actual k-means algorithm rather than the idealized one.

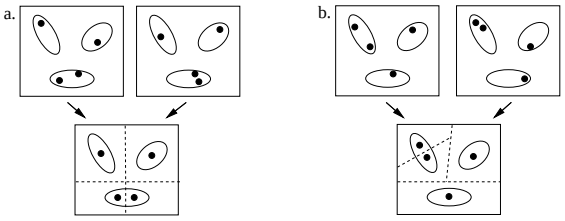

Figure 1: Different initial configurations and the corresponding outcomes of the k-means algorithm. Figure a: the two boxes in the top row depict a data set with three clusters and four initial centers. Both boxes show different realizations of the same initial configuration. As can be seen in the bottom, both initializations lead to the same k-means clustering. Figure b: here the initial configuration is different from the one in Figure a, which leads to a different k-means clustering.

Our analysis theoretically confirms the following intuition. Assume the data set has K well-separated clusters, and assume that k-means is initialized with $K' \geqslant K$ initial centers. We conjecture that when there is at least one initial center in each of the underlying clusters, then *the initial centers tend to stay in the clusters they had been placed in.*

Consequently, the final clustering result is essentially determined by the *number* of initial centers in each of the true clusters (which we call the *initial configuration*), see Figure 1 for an illustration. In particular if one uses an initialization scheme which has the desired property of placing at least one center in each cluster with high probability, then the following will hold: If $K' = K$, we have one center per cluster, with high probability. The configuration will remain the same during the course of the algorithm. If $K' > K$, different configurations can occur. Since different configurations lead to different clusterings we obtain significantly different final clusterings depending on the random initialization, in other words we observe *instability (w.r.t initialization)* .

Note that our argument does not imply stability or instability for $K' < K$. As we have less initial centers than clusters, for any initialization scheme there will be some clusters with no initial center. In this setting centers do move between clusters, and this cannot be analyzed without looking at the actual positions of the centers. Actually, as can be seen from examples, in this case one can have either stability or instability.

The main point of this chapter is that the arguments above can explain why the parameter k selected by stability based model selection is often the true number of clusters, under the assumption that the data set consists of well separated clusters and one uses an appropriate initialization scheme.

Even though the arguments above are very intuitive, even individual parts of our conjecture turn out to be surprisingly hard. In this chapter we only go a first step towards a complete proof, considering mixtures of Gaussians in one dimension. For a mixture of two Gaussians ($K = 2$) we prove that the k-means algorithm is stable for $K' = 2$ and instable for $K' = 3$. The proof technique is based on our configuration arguments outlined above. We also provide some preliminary results to study the general case, that is when the data space is $\mathbb{R}^d$ and we do not make any parametric assumption on the probability distribution. Then we have a closer look at initialization schemes for k-means, when $K' \geqslant K$. Is there an initialization scheme that will place at least one center in each true cluster w.h.p? Clearly, the naive method of sampling K' centers from the

data set does not satisfy this property except for very small K. We study a standard but not naive initialization scheme and prove that it has the desirable property we were looking for.

Of course there exist numerous other papers which study theoretical properties of the actual k-means algorithm. However, these papers are usually concerned with the *value* of the k-means objective function at the final solution, not with the *position* of the final centers. As far as we know, this work is the first one which analyzes the "regions of attractions" of the different local optima of the actual k-means algorithm and derives results on the stability of the k-means clustering itself.

2. Notation and assumptions

In the following we assume that we are given a set of n data points $X_1, ..., X_n \in \mathbb{R}$ which have been drawn i.i.d. according to some underlying distribution $\mathbb{P}$. For a center vector $c = (c_1, ..., c_{K'})$ with $c_i \in \mathbb{R}^d$ we denote the cluster induced by center c_k with $\mathcal{C}_k(c)$. The number of points in this cluster is denoted $N_k(c)$. The clustering algorithm we study in this chapter is the standard k-means algorithm. We denote the initial centers by $c_1^{<0>}, ..., c_{K'}^{<0>}$ with $c_i^{<0>} \in \mathbb{R}$, and the centers after step t of the algorithm as $c_1^{<t>}, ..., c_{K'}^{<t>}$. By K we denote the true number of clusters, by K' the number of clusters constructed by the k-means algorithm. It attempts to minimize the k-means objective function

$$W_n : \mathbb{R}^{dK'} \to \mathbb{R}, W_n(c_1, ..., c_{K'}) = \frac{1}{2} \sum_{i=1}^{n} \min_{k=1,..,K'} ||c_k - X_i||^2.$$

We now restate the k-means algorithm:

Input: $X_1, ..., X_n \in \mathbb{R}^d$, $K' \in \mathbb{N}$

Initialize the centers $c_1^{<0>}, ..., c_{K'}^{<0>} \in \mathbb{R}^d$

Repeat until convergence:

1. Assign data points to closest centers.

2. Re-adjust cluster means:

$$c_k^{<t+1>} = \frac{1}{N_k(c^{<t>})} \sum_{i:\, X_i \in \mathcal{C}_k(c^{<t>})_k} X_i \tag{9.1}$$

Output: $c = (c_1^{<final>}, ..., c_{K'}^{<final>})$.

Traditionally, the instability of a clustering algorithm is defined as the mean (with respect to the random sampling of data points) minimal matching distance between two clusterings obtained on two different set of data points. For the actual k-means algorithm, a second random process is the random initialization (which has not been taken into account in previous literature). Here we additionally have to take the expectation over the random initialization when computing the stability of an algorithm. In this chapter we will derive qualitative rather than quantitative results on stability, thus we omit more detailed formulas.

In the following we restrict our attention to the simple setting where the underlying distribution is a mixture of Gaussians on $\mathbb{R}$ and we have access to an infinite amount of data from $\mathbb{P}$. In particular, instead of estimating means empirically when calculating the new centers of a k-means step we assume access to the true means. In this case, the update step of the k-means algorithm can be written as

$$c_k^{<t+1>} = \frac{\int_{\mathcal{C}_k(c^{<t>})} x f(x) dx}{\int_{\mathcal{C}_k(c^{<t>})} f(x) dx}$$

where f is the density of the probability distribution $\mathbb{P}$. Results in the finite data case can be derived by the help of concentrations inequalities. However, as this introduces heavy notation and our focus lies on the random initialization rather than the random drawing of data points we skip the details. To further set up notation we denote $\varphi_{\mu,\sigma}$ the pdf of a Gaussian distribution with mean μ and variance σ. We also denote $f(x) = \sum_{k=1}^{K} w_k \varphi_{\mu_k,\sigma}$ where K is the number of Gaussians, the weights w_k are positive and sum to one, the means $\mu_{1:K} = (\mu_1, \ldots, \mu_K)$ are ordered, $\mu_1 \leqslant \ldots \leqslant \mu_K$. The minimum separation between two Gaussians is denoted by $\Delta = \min_k(\mu_{k+1} - \mu_k)$. For the standard normal distribution we denote the pdf as φ and the cdf as Φ.

3. The level sets approach

In this section we want to prove that if we run the k-means algorithm with $K' = 2$ and $K' = 3$ on a mixture of two Gaussians, then the resulting clustering depends exclusively on the initial configuration. More precisely if we initialize the algorithm such that each cluster gets at least one center and the initial centers are "close enough" to the true cluster means, then during the course of the algorithm the initial centers do not leave the cluster they had been placed in. This implies stability for $K' = 2$ since there is only one possible configuration satisfying this constraint. On the other hand for $K' = 3$ we have two possible configurations, and thus instability will occur.

The following function plays an important role in our analysis:

$$H : \mathbb{R}^2 \to \mathbb{R},\ H(x, y) = x\Phi(-x + y) - \varphi(-x + y).$$

Straightforward computations show that for any μ, σ, α and h one has

$$\int_{-\infty}^{h} (x - \mu + \alpha)\varphi_{\mu,\sigma}(x)dx = \sigma H\left(\frac{\alpha}{\sigma}, \frac{h + \alpha - \mu}{\sigma}\right). \tag{9.2}$$

We describe necessary and sufficient conditions to obtain stability results for particular "regions" in terms of the level sets of H.

3.1. Stability in the case of two initial centers. We consider the square $S_a = [\mu_1 - a, \mu_1 + a] \times [\mu_2 - a, \mu_2 + a]$ in $\mathbb{R}^2$. The region S_a is called *a stable region* if

$$c^{<0>} \in S_a \Rightarrow c^{<1>} \in S_a \tag{9.3}$$

PROPOSITION 9.1 (Stable region for $K' = 2$). *Equation (9.3) is true if and only if the following four inequalities are satisfied:*

$$\bullet\ w_1 H\left(\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{a + \Delta}{\sigma}, \frac{\Delta}{2\sigma}\right) \geqslant 0 \tag{9.4}$$

$$\bullet\ w_1 H\left(-\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{-a + \Delta}{\sigma}, \frac{\Delta}{2\sigma}\right) \leqslant 0 \tag{9.5}$$

$$\bullet\ w_1 H\left(\frac{a - \Delta}{\sigma}, -\frac{\Delta}{2\sigma}\right) + w_2 H\left(\frac{a}{\sigma}, \frac{\Delta}{2\sigma}\right) \geqslant 0 \tag{9.6}$$

$$\bullet\ w_1 H\left(\frac{-a - \Delta}{\sigma}, -\frac{\Delta}{2\sigma}\right) + w_2 H\left(-\frac{a}{\sigma}, -\frac{\Delta}{2\sigma}\right) \leqslant 0 \tag{9.7}$$

PROOF. Similar to the proof of Proposition 9.2, see below. □

This proposition gives necessary and sufficient conditions for the stability of k-means in the case $K' = 2$. In the following corollary we show an example of the kind of result we can derive from Proposition 9.1. Note that the parameters a and Δ only appear relative to σ. This allows us to consider an arbitrary σ.

COROLLARY 9.1 (Stability for $K' = 2$). *Assume that* $\min(w_1, w_2) = 0.2$ *and* $\Delta = 7\sigma$. *Assume that we have an initialization scheme satisfying:*

- *with probability at least* $1-\delta$ *we have one initial center within* 2.5σ *of* μ_1 *and one within* 2.5σ *of* μ_2.

Then k-means is stable in the sense that with probability at least $1-\delta$ *it converges to a solution with one center within* 2.5σ *of* μ_1 *and one within* 2.5σ *of* μ_2.

PROOF. We simply check numerically that for $a = 2.5\sigma, \Delta = 7\sigma$ and $w_1 = 0.2$ (we also check $w_2 = 0.2$) Equations (9.4) - (9.7) are true. Then by Proposition 9.1 we know that S_a is a stable region which implies the result. □

3.2. Instability in the case of 3 centers. The case of 3 centers gets more intricate. Consider the prism $T_{a,b,\varepsilon}$ and its symmetric version $sym(T_{a,b,\varepsilon})$ in $\mathbb{R}^3$:

$$\begin{aligned} T_{a,b,\varepsilon} =& \{c \in \mathbb{R}^3 : c_1 \leqslant c_2 \leqslant c_3, \\ & c \in [\mu_1 - a, \mu_1 + a - \varepsilon] \times [\mu_1 - a + \varepsilon, \mu_1 + a] \times [\mu_2 - b, \mu_2 + b]\} \\ sym(T_{a,b,\varepsilon}) =& \{c \in \mathbb{R}^3 : c_1 \leqslant c_2 \leqslant c_3, \\ & c \in [\mu_1 - b, \mu_1 + b] \times [\mu_2 - a, \mu_2 + a - \varepsilon] \times [\mu_2 - a + \varepsilon, \mu_2 + a]\}. \end{aligned}$$

If we have an initialization scheme such that each cluster gets at least one center and the initial centers are close enough to the true cluster means, then we initialize either in $T_{a,b,\varepsilon}$ or $sym(T_{a,b,\varepsilon})$. Thus, if these regions are stable in the following sense:

$$c^{<0>} \in T_{a,b,\varepsilon} \Rightarrow c^{<1>} \in T_{a,b,\varepsilon} \tag{9.8}$$

then the global k-means algorithm will be instable, leading either to a clustering in $T_{a,b,\varepsilon}$ or $sym(T_{a,b,\varepsilon})$. Expressed in the terms used in the introduction, the algorithm will be initialized with different configurations and thus be instable.

PROPOSITION 9.2 (Stable region for $K' = 3$). *Equation (9.8) is true if and only if all the following inequalities are satisfied:*

$$\bullet w_1 H\left(\frac{a}{\sigma}, \frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a+\Delta}{\sigma}, \frac{\varepsilon}{2\sigma}\right) \geqslant 0 \tag{9.9}$$

$$\bullet w_1 H\left(\frac{-a+\varepsilon}{\sigma}, \frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta+\varepsilon}{\sigma}, \frac{\varepsilon}{2\sigma}\right) \leqslant 0 \tag{9.10}$$

$$\begin{aligned} &\bullet w_1 H\left(\frac{a-\varepsilon}{\sigma}, \frac{a-b+\Delta-\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a-\varepsilon+\Delta}{\sigma}, \frac{a-b+\Delta-\varepsilon}{2\sigma}\right) \\ &\geqslant w_1 H\left(\frac{a-\varepsilon}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{a-\varepsilon+\Delta}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) \end{aligned} \tag{9.11}$$

$$\begin{aligned} &\bullet w_1 H\left(-\frac{a}{\sigma}, \frac{b-a+\Delta}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta}{\sigma}, \frac{b-a+\Delta}{2\sigma}\right) \\ &\leqslant w_1 H\left(-\frac{a}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{-a+\Delta}{\sigma}, -\frac{\varepsilon}{2\sigma}\right) \end{aligned} \tag{9.12}$$

$$\begin{aligned} &\bullet w_1 H\left(\frac{b-\Delta}{\sigma}, \frac{b-a-\Delta+\varepsilon}{2\sigma}\right) + w_2 H\left(\frac{b-\Delta}{\sigma}, \frac{b-a-\Delta+\varepsilon}{2\sigma}\right) \\ &\leqslant b/\sigma - w_1\Delta/\sigma \end{aligned} \tag{9.13}$$

$$\begin{aligned} &\bullet w_1 H\left(\frac{-b-\Delta}{\sigma}, \frac{a-b-\Delta}{2\sigma}\right) + w_2 H\left(-\frac{b}{\sigma}, \frac{a-b-\Delta}{2\sigma}\right) \\ &\geqslant -b/\sigma - w_1\Delta/\sigma \end{aligned} \tag{9.14}$$

PROOF. *(Sketch)* Let $c^{<0>} \in T_{a,b,\varepsilon}$. Note that the k-means algorithm in one dimension does not change the orders of centers, hence $c_1^{<1>} \leqslant c_2^{<1>} \leqslant c_1^{<3>}$. By the definition of $T_{a,b,\varepsilon}$, to prove that after the first step of k-means the centers $c^{<1>}$ are still in $T_{a,b,\varepsilon}$ we have to check six constraints. Due to space constraints, we only show how to prove that the first constraint $c_1^{<1>} \geqslant \mu_1 - a$ is equivalent to Equation (9.9). The other conditions can be treated similarly.

The update step of the k-means algorithm on the underlying distribution readjusts the centers to the actual cluster means:

$$c_1^{<1>} = \frac{1}{\int_{-\infty}^{\frac{c_1^{<0>}+c_2^{<0>}}{2}} f(x)} \int_{-\infty}^{\frac{c_1^{<0>}+c_2^{<0>}}{2}} x f(x).$$

Thus, $c_1^{<1>} \geqslant \mu_1 - a$ is equivalent to

$$\int_{-\infty}^{\frac{c_1^{<0>}+c_2^{<0>}}{2}} (x - \mu_1 + a) f(x) \geqslant 0.$$

Moreover, the function $h \mapsto \int_{-\infty}^{h} (x - \mu_1 + a) f(x)$ is nondecreasing for $h \in [\mu_1 - a, +\infty)$. Since $c^{<0>} \in T_{a,b,\varepsilon}$ we know that $(c_1^{<0>} + c_2^{<0>})/2 \geqslant \mu_1 - a + \varepsilon/2$ and thus the statement $\forall c^{<0>} \in T_{a,b,\varepsilon}, c_1^1 \geqslant \mu_1 - a$ is equivalent to

$$\int_{-\infty}^{\mu_1 - a + \varepsilon/2} (x - \mu_1 + a) f(x) \geqslant 0.$$

We can now apply Eq. (9.2) with the following decomposition to get Eq. (9.9):

$$\begin{aligned} &\int_{-\infty}^{\mu_1 - a + \varepsilon/2} (x - \mu_1 + a) f(x) \\ &= w_1 \int_{-\infty}^{\mu_1 - a + \varepsilon/2} (x - \mu_1 + a) \varphi_{\mu_1,\sigma} + w_2 \int_{-\infty}^{\mu_1 - a + \varepsilon/2} (x - \mu_2 + \Delta + a) \varphi_{\mu_2,\sigma}. \end{aligned}$$

□

A simple symmetry argument allows us to treat the stability of the symmetric prism.

PROPOSITION 9.3. *If $T_{a,b,\varepsilon}$ is stable for the pdf $f(x) = w_1\varphi_{\mu_1,\sigma} + w_2\varphi_{\mu_2,\sigma}$ and $\widetilde{f}(x) = w_2\varphi_{\mu_1,\sigma} + w_1\varphi_{\mu_2,\sigma}$, then the same holds for $sym(T_{a,b,\varepsilon})$.*

PROOF. The k-means algorithm is invariant with respect to translation of the real axis as well as to changes in its orientation. Hence if $T_{a,b,\varepsilon}$ is stable under f (resp. $\widetilde{f}$), so is $sym(T_{a,b,\varepsilon})$ under $\widetilde{f}(x) = w_2\varphi_{\mu_1,\sigma} + w_1\varphi_{\mu_2,\sigma}$ (resp. f). □

COROLLARY 9.2 (Instability for $K' = 3$). *Assume that $\min(w_1, w_2) = 0.2$ and $\Delta = 14.5\sigma$. Assume that we have an initialization scheme satisfying:*

- *with probability at least $(1-\delta)/2$ we have 2 initial centers within 2.5σ of μ_1 and 1 initial center within 2.5σ of μ_2*

- *with probability at least* $(1-\delta)/2$ *we have 1 initial centers within* 2.5σ *of* μ_1 *and 2 initial centers within* 2.5σ *of* μ_2

Then k-means is instable: with probability $(1-\delta)/2$ *it will converge to a solution with two centers within* 3.5σ *of* μ_1 *and with probability* $(1-\delta)/2$ *to a solution with two centers within* 3.5σ *of* μ_2.

PROOF. We simply check numerically that for $a = 3.5\sigma$, $b = 2.5\sigma$, $\varepsilon = \sigma$, $\Delta = 14.5\sigma$ and $w_1 = 0.2$ (we also check $w_2 = 0.2$) Equations (9.9) - (9.14) are true. Then by Proposition 9.2 and Proposition 9.3 we know that $T_{3.5\sigma,2.5\sigma,\sigma}$ and its symmetric $sym(T_{3.5\sigma,2.5\sigma,\sigma})$ are stable regions which implies the result. □

4. Towards more general results: the geometry of the solution space of k-means

In the section above we proved by a level set approach that in a very simple setting, if we initialize the k-means algorithm "close enough" to the true cluster centers, then the initial centers do not move between clusters. However we would like to obtain this result in a more general setting. We believe that to achieve this goal in a systematic way one has to understand the structure of the solution space of k-means. We identify the solution space with the space $\mathbb{R}^{dK'}$ by representing a set of K' centers $c_1, ..., c_{K'} \in \mathbb{R}^d$ as a point c in the space $\mathbb{R}^{dK'}$. Our goal in this section is to understand the "shape" of the k-means objective function on this space. Secondly, we want to understand how the k-means algorithm operates on this space. That is, what can we say about the "trajectory" of the k-means algorithm from the initial point to the final solution? For simplicity, we state some of the results in this section only for the case where the data space is one dimensional. They also hold in $\mathbb{R}^d$, but are more nasty to write up.

First of all, we want to compute the derivatives of W_n with respect to the individual centers.

PROPOSITION 9.4 (Derivatives of k-means). *Given a finite data set $X_1, ..., X_n \in \mathbb{R}$. For $k, l \in \{1, ..., K'\}$ and $i \in \{1, ..., n\}$ consider the hyperplane in $\mathbb{R}^{K'}$ which is defined by*

$$H_{k,l,i} := \{c \in \mathbb{R}^{K'} : X_i = (c_k + c_l)/2\}.$$

Define the set $H := \cup_{k,l=1}^{K'} \cup_{i=1}^{n} H_{k,l,i}$. Then we have:

(1) *W_n is differentiable on $\mathbb{R}^{K'} \setminus H$ with partial derivatives*

$$\frac{\partial W_n(c)}{\partial c_k} = \sum_{i:\, X_i \in \mathcal{C}_k} (c_k - X_i).$$

(2) *The second partial derivatives of W_n on $\mathbb{R}^{K'} \setminus H$ are*

$$\frac{\partial W_n(c)}{\partial c_k \partial c_l} = 0 \qquad \text{and} \qquad \frac{\partial W_n(c)}{\partial c_k \partial c_k} = N_k.$$

(3) *The third derivatives of W_n on $\mathbb{R}^{K'} \setminus H$ all vanish.*

PROOF. First of all, note that the sets $H_{k,l,i}$ contain the center vectors for which there exists a data point X_i which lies on the boundary of two centers c_k and c_l. Now let us look at the first derivative. We compute it by foot:

$$\frac{\partial W_n(c)}{\partial c_k} = \lim_{h \to 0} \frac{1}{h}(W_n(c_1, ..., c_K) - W_n(c_1, ..., c_k + h, ..., c_K))$$

When $c \notin H$ we know that no data point lies on the boundary between two cluster centers. Thus, if h is small enough, the assignment of data points to cluster centers does not change if we replace c_k by $c_k + h$. With this property, the expression above is trivial to compute and yields the first derivative, the other derivatives follow similarly. □

A straightforward consequence is as follows:

PROPOSITION 9.5 (k-means does Newton iterations). *The update steps performed by the k-means algorithms are exactly the same as update steps by a Newton optimization.*

PROOF. This proposition follows directly from Proposition 9.4, the definition of the Newton iteration on W_n and the definition of the k-means update step. This fact has also been stated (less rigorously and without proof) in Bottou and Bengio [1995]. □

Together, the two propositions show an interesting picture. We have seen in Proposition 9.4 that the k-means objective function W_n is differentiable on $\mathbb{R}^{K'} \setminus H$. This means that the space $\mathbb{R}^{K'}$ is separated into many cells with hyperplane boundaries $H_{k,l,i}$. By construction, the cells are convex (as they are intersections of half-spaces). Our finding means that each data set $X_1, ..., X_n$ induces a partitioning of this solution space into convex cells. To avoid confusion, at this point we would like to stress again that we are not looking at a fixed clustering solution on the data space (which can be described by cells with hyperplane boundaries, too), but at the space of all center vectors c. It is easy to see that all centers c within one cell correspond to exactly one clustering of the data points. As it is well known that the k-means algorithm never visits a clustering twice, we can conclude that each cell is visited at most once by the algorithm.Within each cell, W_n is quadratic (as the third derivatives vanish). Moreover, we know that k-means behaves as the Newton iteration. On a quadratic function, the Newton optimization jumps in one step to the minimum of the function. This means that if k-means enters a cell that contains a local optimum of the k-means objective function, then the next step of k-means jumps to this local optimum and stops.

Now let us look more closely at the trajectories of the k-means algorithm. The paper by Zhang et al. [2008] inspired us to derive the following property.

PROPOSITION 9.6 (Trajectories of k-means). *Let $c^{<t>}$ and $c^{<t+1>}$ be two consecutive solutions visited by the k-means algorithm. Consider the line connecting those two solutions in $\mathbb{R}^{K'}$, and let $c^\alpha = (1-\alpha)c^{<t>} + \alpha c^{<t+1>}$ be a point on this line (for some $\alpha \in [0,1]$). Then $W_n(c^\alpha) \leqslant W_n(c^{<t>})$.*

PROOF. The following inequalities hold true:

$$\begin{aligned}
W_n(c^\alpha) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^\alpha)} ||X_i - c_k^\alpha||^2 \\
&\leqslant \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^t)} ||X_i - c_k^\alpha||^2 \\
&\leqslant \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{C}_k(c^t)} \alpha||X_i - c_k^t||^2 + (1-\alpha)||X_i - c_k^{t+1}||^2 \\
&\leqslant \alpha W_n(c^t) + (1-\alpha)W_n(c^{t+1})
\end{aligned}$$

For the first and third inequality we used the fact that assigning points in $\mathcal{C}_k(c)$ to the center c_k is the best thing to do to minimize W_n. For the second inequality we used that $x \to ||x||^2$ is convex. The proof is concluded by noting that $W_n(c^{<t>}) \leqslant W_n(c^{<t+1>})$. □

We believe that the properties of the k-means objective function and the algorithm are the key to prove more general stability results. However, there is still an important piece missing, as we are going to explain now. Since k-means performs Newton iterations on W_n, one could expect

to get information on the trajectories in the configuration space by using a Taylor expansion of W_n. However, as we have seen above, each step of the k-means algorithm crosses one of the hyperplanes $H_{k,l,i}$ on which W_n is non-differentiable. Hence, a direct Taylor expansion approach on W_n cannot work. On the other hand, surprisingly one can prove that the limit objective function $W := \lim \frac{1}{n} W_n$ is almost surely a continuously differentiable function on $\mathbb{R}^{K'}$ (we omit the proof). Thus one may hope that one could first study the behavior of the algorithm for W, and then apply concentration inequalities to carry over the results to W_n. Unfortunately, here we face another problem: one can prove that in the limit case, a step of the k-means algorithm is *not* a Newton iteration on W.

Proposition 9.6 directly evokes a scheme to design stable regions. Assume that we can find two regions $A \subset B \subset \mathbb{R}^{K'}$ of full rank and such that

$$\max_{x \in \partial A} W_n(x) \leqslant \min_{x \in \partial B} W_n(x). \tag{9.15}$$

Then, if we initialize in A we know that we will converge to a configuration in B. This approach sounds very promising. However, we found that it was impossible to satisfy both Equation (9.15) and the constraint that A has to be "big enough" so that we initialize in A with high probability.

Finally, we would like to elaborate on a few more complications towards more general results:

- On a high level, we want to prove that if K' is slightly larger than the true K, then k-means is instable. On the other hand, if K' gets close to the number n of data points, we trivially have stability again. Hence, there is some kind of "turning point" where the algorithm is most instable. It will be quite a challenge to work out how to determine this turning point.
- Moreover, even if we have so many data points that the above problem is unlikely to occur, our analysis breaks down if K' gets too large. The reason is that if K' is much bigger than K, then we cannot guarantee any more that initial centers will be in stable regions. Just the opposite will happen: at some point we will have outliers as initial centers, and then the behavior of the algorithm becomes rather unpredictable.
- Finally, consider the case of $K' < K$. As we have already mentioned in the introduction, in this case it is not necessarily the case that different initial configurations lead to different clusterings. Hence, a general statement on (in)stability is not possible in this case. This also means that the tempting conjecture "the true K has minimal stability" is not necessarily true.

5. An initialization algorithm and its analysis

We have seen that one can prove results on clustering stability for k-means if we use a "good" initialization scheme which tends to place initial centers in different Gaussians. We now show that an established initialization algorithm, the PRUNED MINDIAM initialization described in Figure 2 has this property, i.e it has the effect of placing the initial centroids in disjoint, bounded neighborhoods of the means $\mu_{1:K}$. This often rediscovered algorithm is credited to Hochbaum and Shmoys [1985]. In Dasgupta and Schulman [2007] it was analyzed it in the context of the EM algorithm. Later N.Srebro et al. [2006] used it in experimental evaluations of EM, and it was found to have a significant advantage w.r.t more naive initialization methods in some cases. While this and other initializations have been extensively studied in conjunction with EM, we are not aware of any studies of PRUNED MINDIAM for k-means.

We make three necessary conceptual assumptions. Firstly to ensure that K is well-defined we assume that the mixture weights are bounded below by a known weight w_{min}. **Assumptions.**

Algorithm PRUNED MINDIAM

Input: w_{min}, number of centers K'

(1) Initialize with L random points $c_{1:L}^{<0>}$, L computed by (9.19)

(2) Run one step of k-means, that is

  (a) To each center $c_j^{<0>}$ assign region $\mathcal{C}_j^0$, $j = 1 : L$

  (b) Calculate $c_{1:L}^{<1>}$ as the centers of mass of regions $\mathcal{C}_{1:L}^0$

(3) Remove all centers $c_j^{<1>}$ for which $P[\mathcal{C}_j^1] \leqslant p_0$, where p_0 is given by (9.19). We are left with $c_{j'}^{<1>}$, $j' = 1 : L'$.

(4) Choose K' of the remaining centers by the MINDIAM heuristic

  (a) Select one center at random.

  (b) Repeat until K' centroids are selected:
  Select the centroid $c_q^{<1>}$ that maximizes the minimum distance to the already selected centroids.

Output: the K' selected centroids $c_k^{<1>}$, $k = 1 : K'$

Figure 2: The PRUNED MINDIAM initialization

$w_k \geqslant w_{min}$ *for all k.* We also require to know a lower bound Δ and an upper bound Δ_{max} on the separation between two Gaussians, and we assume that these separations are "sufficiently large". In addition, later we shall make several technical assumptions related to a parameter τ used in the proofs, which also amount to conditions on the separation. These assumptions shall be made precise later.

THEOREM 9.1 (PRUNED MINDIAM Initialization). *Let $f = \sum_1^K w_k \varphi_{\mu_k,1}$ be a mixture of K Gaussians with centers $\mu_{1:K}$, $\mu_k \leqslant \mu_{k+1}$, and unit variance. Let $\tau \in (0, 0.5)$, $\delta_{miss} > 0$, δ_{impure} defined in Proposition 9.8. If we run Algorithm* PRUNED MINDIAM *with any $2 \leqslant K' \leqslant 1/w_{min}$, then, subject to Assumptions 1, 2, 3, 4, 5 (specified later), with probability $1 - 2\delta_{miss} - \delta_{impure}$ over the initialization there exist K disjoint intervals $\widetilde{A}_k$, specified in Section 5.4, one for each true mean μ_k, so that all K' centers $c_{k'}^{<1>}$ are contained in $\bigcup_k \widetilde{A}_k$ and*

(9.16) *if $K' = K$, each $\widetilde{A}_k$ will contain exactly one center $c_{k'}^{<1>}$,*

(9.17) *if $K' < K$, each $\widetilde{A}_k$ will contain at most one center $c_{k'}^{<1>}$,*

(9.18) *if $K' > K$, each $\widetilde{A}_k$ will contain at least one center $c_{k'}^{<1>}$.*

The idea to prove this result is to show that the following statements hold with high probability. By selecting L preliminary centers in step 1 of PRUNED MINDIAM, each of the Gaussians obtains at least one center (Section 5.1). After steps 2a, 2b we obtain "large" clusters (mass $> p_0$) and "small" ones (mass $\leqslant p_0$). A cluster can also be "pure" (respectively "impure") if most of its mass comes from a single Gaussian (respectively from several Gaussians). Step 3 removes all "small" cluster centers, but (and this is a crucial step of our argument) w.h.p it will also remove all "impure" cluster centers (Section 5.2). The remaining clusters are "pure" and "large"; we show (Section 5.3) that each of their centers is reasonably close to some Gaussian mean μ_k. Hence, if the Gaussians are well separated, the selection of final centers $c_q^{<1>}$ in step 4 "cycles through different Gaussians" before visiting a particular Gaussian for the second time (Section 5.4). The rest of this section outlines these steps in more details.

5.1. Step 1 of PRUNED MINDIAM. Picking the initial centroids $c^{<0>}$. We need to pick a number of initial centers L large enough that each Gaussian has at least 1 center w.h.p. We

formalize this here and find a value for L that ensures the probability of this event is at least $1-\delta_{miss}$, where δ_{miss} is a tolerance of our choice. Another event that must be avoided for a "good" initialization is that all centroids $c_j^{<0>}$ belonging to a Gaussian end up with initial clusters $\mathcal{C}_j^0$ that have probability less than p_0. If this happens, then after thresholding, the respective Gaussian is left with no representative centroid, i.e it is "missed". We set the tolerance for this event to $\delta_{thresh} = \delta_{miss}$. Let $t = 2\Phi(-\Delta/2)$ the *tail probability* of a cluster and A_k the symmetric neighborhood of μ_k that has $\varphi_{\mu_k,1}(A_k) = 1 - t$.

PROPOSITION 9.7. *If we choose*

$$L \geqslant \left(\ln \frac{1}{\delta_{miss} w_{min}}\right) \Big/ \left((1-t)w_{min}\right) \quad \textit{and} \quad p_0 = \frac{1}{eL} \tag{9.19}$$

then the probability over all random samplings of centroids $c_{1:L}^{<0>}$ that at least one centroid $c_j^{<0>}$ with assigned mass $P[\mathcal{C}_j^0] \geqslant p_0$ can be found in each A_k, $k = 1 : K$, is greater or equal to $1 - 2\delta_{miss}$.

The proof of this result is complicated but standard fare (e.g. Chernoff bounds) and is therefore omitted.

After steps 1, 2a and 2b of PRUNED MINDIAM are performed, we obtain centers $c_{1:L}^{<1>}$ situated at the centers of mass of their respective clusters $\mathcal{C}_{1:L}^1$. Removing the centers of small clusters follows. We now describe a beneficial effect of this step.

5.2. Step 3 of PRUNED MINDIAM. Thresholding removes impure clusters. We introduce the concept of *purity* of a cluster, which is related to the ratio of points from a certain Gaussian w.r.t to the total probability mass of the cluster. Denote P_k the probability distribution induced by the k-th Gaussian $\varphi_{\mu_k,1}$.

DEFINITION 9.1. *A cluster $\mathcal{C}$ is* $(1-\tau)$-pure *if most of its points come from a single Gaussian, i.e if $w_k P_k[\mathcal{C}] \geqslant (1-\tau)P[\mathcal{C}]$, with $\tau < 1/2$ being a positive constant. A cluster which is not $(1-\tau)$-pure is* τ-impure *(or simply* impure*).*

The values of τ that we consider useful are of the order $0.001 - 0.02$ and, as it will appear shortly, $\tau < w_{min}/2$. The purity of a cluster helps in the following way: if a cluster is pure, then it can be "tied" to one of the Gaussians. Moreover, its properties (like center of mass) will be dictated by the Gaussian to which it is tied, with the other Gaussians' influence being limited; Section 5.3 exploits this idea.

But there will also be clusters that are impure, and so they cannot be tied to any Gaussian. Their properties will be harder to analyze, and one expects their behavior to be less predictable. Luckily, impure clusters are very likely small. As we show now, the chance of having an impure cluster with mass larger than p_0 is bounded by a δ_{impure} which we are willing to tolerate.

Because of limited space, we leave out the long and complex rigourous proofs of this result, and give here just the main ideas. Let $\mathcal{C}_j = [z_1, z_2]$ be a τ-impure cluster, with $P[\mathcal{C}_j] \geqslant p_0$, c_j the centroid that generates $\mathcal{C}_j$ (not necessarily at its center of mass) and c_{j-1}, c_{j+1} the centroids of the adjacent clusters (not necessarily centers of mass). As one can show, even though an impure cluster contains some probability mass from each Gaussian, in most of this section we only need consider the two Gaussians which are direct neighbors of $\mathcal{C}$. Let us denote the parameters of these (consecutive) Gaussians by $\mu_{1,2}, w_{1,2}$.

For the purpose of the proof, we are looking here at the situation after step 2a, thus the centroids $c_{j-1,j,j+1}$ should be $c_{j-1,j,j+1}^{<0>}$, but we renounce this convention temporarily to keep the notation light. We want to bound the probability of cluster $\mathcal{C}_j$ being impure and large. Note that

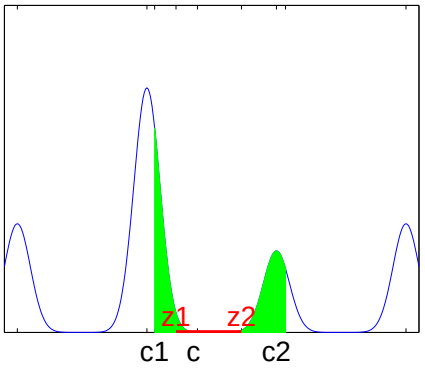

Figure 3: Concrete example of a large impure cluster $[z1, z2]$; $c1, c, c2$ represent the consecutive cluster centers $c_{j-1}^{<0>}, c_j^{<0>}, c_{j+1}^{<0>}$. We demonstrate that if $P[z1, z2] > p_0$ then the interval $[c1, c2]$ (which is twice its length) must have mass $> p_1 >> p_0$. If L is large enough, having such a large interval contain a single c_j is improbable. Numerical values: mixture with $\Delta = 10, w_{min} = 0.15$, impurity $\tau([z_1, z_2]) = 0.07$, $P[z1, z2] = 0.097$, $P[c1, c2] = 0.24$; using $\delta_{miss} = 0.02, \tau = 0.015$ one gets $L = 38, p_0 = 0.095 < P[z1, z2]$, $p_1 = 0.0105 < P[c1, c2]$, $\delta_{impure} = 0.016 >> (1 - P[c1, c2])^{L-1} = 0.00003$

Step 2b of the PRUNED MINDIAM does not affect either of these properties, as it only acts on the centers.

A simple observation is the following. Since $z_1 = \frac{c_{j-1}+c_j}{2}$ and $z_2 = \frac{c_{j+1}+c_j}{2}$ we have $c_{j+1} - c_{j-1} = 2(z_2 - z_1) = 2\Delta z$. The idea is to show that if an impure region has probability larger than p_0, then the interval $[c_{j-1}, c_{j+1}]$ has probability at least p_1, significantly larger than p_0. On the other hand, the probability of sampling from P a single center $\mathcal{C}_j$ out of a total of L in an interval of length $2\Delta z$ is $P[c_{j-1}, c_{j+1}](1-P[c_{j-1}, c_{j+1}])^{L-1} < (1-p_1)^{L-1}$. If p_1 and L are large enough, then $(1 - p_1)^{L-1} \stackrel{def}{=} \delta_{impure}$ will be vanishingly small. We proceed in two steps: first we find the minimum length Δz_0 of a cluster $\mathcal{C}_j$ which is impure and large. Then, we find a lower bound p_1 on the probabability of any interval $[c,\ c + 2\Delta z_0]$ under the mixture distribution. The following assumption ensures that the purity $1 - \tau$ is attainable in each Gaussian. **Assumptions.**

Let $\gamma_{k,k'}(x) = \frac{w_{k'}\varphi_{\mu_{k'},1}(x)}{w_k\varphi_{\mu_k,1}(x)}$ *(a local purity measure). Then*

$$\sum_{k' \neq k} \gamma_{k,k'} \left(\Phi^{-1} \left(\frac{1}{2} + \frac{(1-\tau)p_0}{2w_{min}} \right) \right) \leqslant \frac{\tau}{1-\tau}.$$

The next assumption ensures that $\Delta z_0 > 0$, i.e it is an informative bound. **Assumptions.** $d\left(\frac{\tau p_0}{w_{min}}\right) < \frac{1}{2}\Delta$.

PROPOSITION 9.8 (Impure clusters are small w.h.p). *Let* w_1, w_2 *be the mixture weights of two consecutive Gaussians and define* $\Delta z_0 = \Delta - d\left(\frac{\tau p_0}{w_1}\right) - d\left(\frac{\tau p_0}{w_2}\right)$,

$$p_1 = w_1 \Phi \left(\frac{\Delta - 2\Delta z_0}{2} - \frac{\ln \frac{w_1}{w_2}}{\Delta - 2\Delta z_0} \right) + w_2 \Phi \left(\frac{\Delta - 2\Delta z_0}{2} - \frac{\ln \frac{w_2}{w_1}}{\Delta - 2\Delta z_0} \right)$$

and $\delta_{impure} = (1-p_1)^{L-1}$. *Let* $\mathcal{C}_j^0, j = 1, \ldots, L$ *be the regions associated with* $c_{1:L}^{<0>}$ *after step 2a of the* PRUNED MINDIAM *algorithm. If assumptions 5,5.2,5.2 hold, then the probability that there exists* $j \in \{1, \ldots, L\}$ *so that* $P[\mathcal{C}_j^0] \geqslant p_0$ *and* $w_1 P_1[\mathcal{C}_j^0] \geqslant \tau P[\mathcal{C}_j^0],\ w_2 P_2[\mathcal{C}_j^0] \geqslant \tau P[\mathcal{C}_j^0]$ *is at most* δ_{impure}. *This probability is over the random initialization of the centroids* $c_{1:L}^{<0>}$.

To apply this proposition without knowing the values of w_1, w_2 one needs to minimize the bound p_1 over the range $w_1, w_2 > w_{min},\ w_2 + w_1 \leqslant 1 - (K-2)w_{min}$. This minimum can be obtained numerically if the other quantities are known.

We also stress that because of the two-step approach, first minimizing Δz_0, then $P[c, c + 2\Delta z_0]$, the bound δ_{impure} obtained is not tight and could be significantly improved.

5.3. The $(1-\tau)$-pure cluster. Now we focus on the clusters that have $P[\mathcal{C}] > p_0$ and are $(1-\tau)$-pure. By Proposition 9.8, w.h.p their centroids are the only ones which survive the thresholding in step 3 of the PRUNED MINDIAM algorithm. In this section we will find bounds on the distance $|c_j^{<1>} - \mu_k|$ between $\mathcal{C}_j$'s center of mass and the mean of "its" Gaussian.

We start by listing some useful properties of the standard Gaussian. Denote by $r(x)$ the center of mass of $[x, \infty)$ under the truncated standard Gaussian, and by $d(t)$ the solution of $1 - \Phi(d) = t$, with $0 < t < 1$. Intuitively, $d(t)$ is the cutoff location for a tail probability of t. Note that any interval whose probability under the standard normal exceeds t must intersect $[-d(t), d(t)]$. Let $a > 0$ (in the following a as to be thought as a small positive constant).

PROPOSITION 9.9. *(i)* $r(x)$ *is convex, positive and increasing for* $x \geqslant 0$ *(ii) For* $w \in [2a, \infty)$ *the function* $d(a/w)$ *is convex, positive and increasing w.r.t* w, *and* $r(d(a/w))$ *is also convex, positive and increasing.*

PROPOSITION 9.10. *Let* $\mathcal{C} = [z_1, z_2]$ *be an interval (with* z_1, z_2 *possibly infinite),* c *its center of mass under the normal distribution* $\varphi_{\mu,1}$ *and* $P[\mathcal{C}]$ *its probability under the same distribution. If* $1/2 \geqslant P[\mathcal{C}] \geqslant p$, *then* $|c - \mu| \leqslant r(d(p))$ *and* $\min\{|z_1 - \mu|, |z_2 - \mu|\} \leqslant d(p) = -\Phi^{-1}(p)$.

The proofs are straightforward and omitted. Define now $w_{max} = 1 - (K-1)w_{min}$ the maximum possible cluster size in the mixture and

$$R(w) = r\left[-\Phi^{-1}\left(\frac{(1-\tau)p_0}{w}\right)\right], \quad \widetilde{R}(w_1, w_2) = -\Phi^{-1}\left[\frac{\tau w_1}{(1-\tau)w_2} + \Phi(d(\frac{(1-\tau)p_0}{w_1} - \Delta)\right]$$

In the next proposition, we will want to assume that $\widetilde{R} \geqslant 0$. The following assumption is sufficient for this purpose. **Assumptions.** $\frac{\tau}{w_{min}} \leqslant \frac{1}{2} - \Phi(-\Delta/2)$

PROPOSITION 9.11 (The $(1-\tau)$-pure cluster). *Let cluster* $\mathcal{C} = [z_1, z_2]$ *with* $z_2 > \mu_k$, $P[\mathcal{C}] \geqslant p_0$ *and* $w_k P_k[\mathcal{C}] \geqslant (1-\tau)P[\mathcal{C}]$ *for some* k, *with* τ *satisfying Assumptions 5.2 and 5.3. Let* c, c_k *denote the center of mass of* $\mathcal{C}$ *under* P, P_k *respectively. Then*

$$|c_k - \mu_k| \leqslant R(w_k) \tag{9.20}$$

and, whenever $k < K$

$$z_2 - \mu_k \leqslant -\widetilde{R}(w_k, w_{k+1}) \leqslant -\widetilde{R}(w_{max}, w_{min}) \tag{9.21}$$

PROPOSITION 9.12 (Corollary). *If* $c_k > \mu_k$ *and* $k < K$ *then*

$$c - \mu_k \leqslant (1-\tau)R(w_k) + \tau(\Delta - \widetilde{R}(w_k, w_{k+1})) \tag{9.22}$$

$$\leqslant (1-\tau)R(w_{max}) + \tau(\Delta - \widetilde{R}(w_{max}, w_{min})) \tag{9.23}$$

$$\leqslant (1-\tau)R(w_{max}) + \tau\Delta \tag{9.24}$$

else

$$\mu_k - c \leqslant R(w_k) \;\leqslant\; R(w_{max}) \quad c - \mu_k \leqslant \tau(\Delta - \widetilde{R}(w_k, w_{k+1})) \tag{9.25}$$

By symmetry, a similar statement involving $\mu_{k-1}, w_{k-1}, \mu_k, w_k$ and c holds when $z_2 > \mu_k$ is replaced by $z_1 < \mu_k$. With it we have essentially shown that an almost pure cluster which is not small cannot be too far from its Gaussian center μ_k.

Proof of Proposition 9.11 (9.20) follows from Proposition 9.10. Now for bounding z_2, in the case $k < K$. Because $(1-\tau)P[\mathcal{C}] \leqslant w_k$ (the contribution of Gaussian k to cluster $\mathcal{C}$ cannot exceed all of w_k) we have $P_{k+1}[C] \leqslant \frac{\tau P[\mathcal{C}]}{w_{k+1}} \leqslant \frac{\tau w_k}{(1-\tau)w_{k+1}}$ and $P_{k+1}[C] = \Phi(z_2 - \mu_{k+1}) - \Phi(z_1 - \mu_{k+1}) \geqslant \Phi(z_2 - \mu_{k+1}) - \Phi(c_1 - \mu_{k+1})$ from which the first inequality in (9.21) follows. The function $\widetilde{R}$ is increasing with w_k when w_{k+1} constant or $w_{k+1} = \text{constant} - w_1$, which gives the second bound. □

Proof of the corollary First note that we can safely assume $z_1 \geqslant \mu_k$. If the result holds for this case, then it is easy to see that having $z_1 < \mu_k$ only brings the center of mass c closer to μ_k.

$$c = \frac{w_k P_k[C] c_k + \sum_{k' \neq k} w_{k'} P_{k'}[C] c_{k'}}{P[\mathcal{C}]} \leqslant (1-\tau)c_k + \tau z_2 \tag{9.26}$$

Now (9.22,9.23) follow from Proposition 9.11. For (9.24) Assumption 5.3 assures that $\widetilde{R} \geqslant 0$. As a consequence, this bound is convex in w_k. If $k = 1$ and $c_1 \leqslant \mu_1$, or $k = K$ and $c_K > \mu_K$ then the second term in the sum (9.26) pulls c_1 in the direction of μ_1 (respectively c_K in the direction of μ_K) and we can get the tighter bounds (9.25). □

In conclusion, we have shown now that if the unpruned center c "belongs" to Gaussian k, then

$$c \in \widetilde{A}_k(w_k) \;=\; [\,\mu_k - R_\tau^-(w_k),\, \mu_k + R_\tau^+(w_k)\,]$$

whith $R_\tau^-(w_k) = (1-\tau)R(w_k) + \tau(\mu_k - \mu_{k-1})$, $R_\tau^+(w_k) = (1-\tau)R(w_k) + \tau(\mu_{k+1} - \mu_k)$, $R_\tau^-(w_1) = R(w_1)$, and $R_\tau^+(w_K) = R(w_K)$.

5.4. Step 4 of PRUNED MINDIAM**. Selecting the centers by the** MINDIAM **heuristic.** From Section 5.2 we know that w.h.p all centroids unpruned at this stage are $(1-\tau)$ pure. We want to ensure that after the selection in step 4 each Gaussian has at least one $c_j^{<1>}$ near its center. For this, it is sufficient that the regions $\widetilde{A}_k$ are disjoint, i.e

$$(\mu_{k+1} - \mu_k) - (R_\tau^+(w_k) + R_\tau^-(w_{k+1})) > R_\tau^-(w_k) + R_\tau^+(w_k)$$
$$(\mu_{k+1} - \mu_k) - (R_\tau^+(w_k) + R_\tau^-(w_{k+1})) > R_\tau^-(w_{k+1}) + R_\tau^+(w_{k+1})$$

for all k. Replacing $R_\tau^\pm(w_k)$ with their definitions and optimizing over all possible $w_{1:K} \geqslant w_{min}$ and for all $\Delta\mu \leqslant \mu_{k+1} - \mu_k \leqslant \Delta_{max}$ produces

$$\widetilde{A}_k \;=\; [\mu_k \pm (1-\tau)R(w_{max}) \pm \tau\Delta_{max}]$$

and **Assumptions.** $(1-3\tau)\Delta - \tau\Delta_{max} \;>\; [3R(w_{max}) + R(w_{min})](1-\tau)$.

6. Simulations

In this section we test our conjecture in practice and run some simulations to emphasize the different theoretical results of the previous sections. We also investigate whether it is necessary to look at the stability of k-means with respect to the random drawing of the data set. In the following when we refer to randomization we mean with respect to the initialization while the resampling corresponds to the random drawing of the data set.

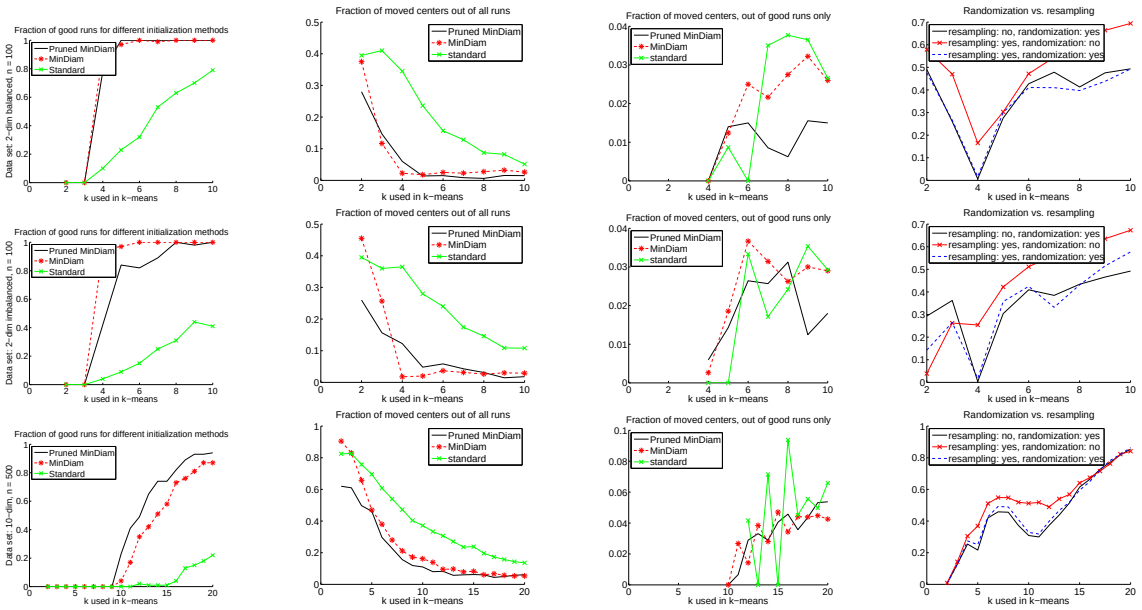

Figure 4: Simulation results. First row: data set "two dim four balanced clusters". Second row: data set "two dim four imbalanced clusters". Third row: data set "ten dim ten clusters" (see text for details).

Setup of the experiments. As distributions we consider mixtures of Gaussians in one, two, and ten dimensions. Each mixture consists of several, reasonably well separated clusters. We report the results on three such data sets:

- "Two dim four balanced clusters": Mixture of four Gaussians in $\mathbb{R}^2$ with means (-3.3), $(0,0)$, $(3,3)$, $(3,-3)$; the covariance matrix of all clusters is diagonal with entries 0.2 and 1 on the diagonal; the mixing coefficients are uniform, that is all clusters have the same weight.
- "Two dim four imbalanced clusters": As above, but with mixing coefficients $0.1, 0.5, 0.3, 0.1$.
- "Ten dim ten clusters": Mixture of ten Gaussians in $\mathbb{R}^{10}$ with means $(i,0,0,...)$ for $i = 1,...,10$. All Gaussians are spherical with variance 0.05 and mixing coefficients are uniform.

As clustering algorithm we use the standard k-means algorithm with the following initializations:

- Standard initialization: randomly pick K' data points.
- MINDIAM initialization, coincides with Step 5 in Fig. 2.
- PRUNED MINDIAMinitialization, as analyzed in Section 5 (see Fig. 2
- Deterministic initialization: K' fixed points sampled from the distribution.

For a range of parameters $K' \in \{2,...,10\}$ we compute the clustering stability by the following protocols:

- Randomization, no resampling: We draw once a data set of $n = 100$ points from the distribution. Then we run the k-means algorithm with different initializations.
- Resampling, no randomization: We fix a set of deterministic starting points (by drawing them once from the underlying distribution). Then we draw 100 data sets of size $n = 100$ from the underlying distribution, and run k-means with the deterministic starting points on these data sets.
- Resampling and randomization: we combine the two previous approaches.

Then we compute the stability with respect to the minimal matching distance between the clusters. Each experiment was repeated 100 times, we always report the mean values over those repetitions.

Note that all experiments were also conducted with different data set sizes ($n = 50, 100, 500$), stability was computed with and without normalization (we used the normalization suggested in Lange et al., 2004), and the k-means algorithm was used with and without restarts. All those variations did not significantly effect the outcome, hence we omit the plots.

Results. First we evaluate the effect of the different initializations. To this end, we count how many initializations were "good initializations" in the sense that each true cluster contains at least one initial center. In all experiments we consistently observe that both the pruned and non-pruned min diameter heuristic already achieve many good runs if K' coincides with K or is only slightly larger than the true K (of course, good runs cannot occur for $K' < K$). The standard random initialization does not achieve the same performance. See Figure 4, first column.

Second, we record how often it was the case that initial cluster centers cross cluster borders. We can see in Figure 4 (second column) that this behavior is strongly correlated with the number of "good initializations". Namely, for initialization methods which achieve a high number of good initializations the fraction of centers which cross cluster borders is very low. Moreover, one can see in the third column of Figure 4 that centers usually do not cross cluster borders if the initialization was a good one. This coincides with our theoretical results.

Finally, we compare the different protocols for computing the stability: using randomization but no resampling, using resampling but no randomization, and using both randomization and resampling, cf. right most plots in Figure 4. In simple data sets, all three protocols have very similar performance, see for example the first row in Figure 4. That is, the stability values computed on the basis of resampling behave very similar to the ones computed on the basis of randomization, and all three methods clearly detect the correct number of clusters. Combining randomization and resampling does not give any advantage. However, on the more difficult data sets (the imbalanced one and the 10-dimensional one), we can see that resampling without randomization performs worse than the two protocols with randomization (second and third row of Figure 4). While the two protocols using randomization have a clear minimum around the correct number of clusters, stability based on resampling alone fails to achieve this. We never observed the opposite effect in any of our simulations (we ran many more experiments than reported here). This shows, as we had hoped, that randomization plays an important role for clustering stability, and in certain settings can achieve better results than resampling alone.

Finally, in the experiments above we ran the k-means algorithm in two modes: with restarts, where the algorithm is started 50 times and only the best solution is kept; and without restarts. The results did not differ much (above we report the results without restarts). This means that in practice, for stability based parameter selection one can save computing time by simply running k-means without restarting it many times (as is usually done in practice). From our theory we had even expected that running k-means without restarts achieves better results than with restarts. We thought that many restarts diminish the effect of exploring local optima, and thus induce more stability than "is there". But the experiments did not corroborate this intuition.

7. Conclusions and outlook

Previous theoretical work on model selection based on the stability of the k-means algorithm has assumed an "ideal k-means algorithm" which always ends in the global optimum of the objective function. The focus was to explain how the random drawing of sample points influences the positions of the final centers and thus the stability of the clustering. This analysis explicitly excluded the question when and how the k-means algorithm ends in different local optima. In particular, this means that these results only have a limited relevance for the actual k-means algorithm as used in practice.

In this chapter we study the actual k-means algorithm. We have shown that the initialization strongly influences the k-means clustering results. We also show that if one uses a "good" initialization scheme, then the k-means algorithm is stable if it is initialized with the correct number of centers, and instable if it is initialized with too many centers. Even though we have only proved these results in a simple setting so far, we are convinced that the same mechanism also holds in a more general setting.

These results are a first step towards explaining why the selection of the number of clusters based on clustering stability is so successful in practice Lange et al. [2004]. From this practical point of view, our results suggest that introducing randomness by the initialization may be sufficient for an effective model selection algorithm. Another aspect highlighted by this work is that the situations $K' < K$ and $K' > K$ may represent two distinct regimes for clustering, that require separate concepts and methods to be analyzed.

The main conceptual insight in the first part of the chapter is the configurations idea described in the beginning. With this idea we indirectly characterize the "regions of attraction" of different local optima of the k-means objective function. To our knowledge, this is the first such characterization in the vast literature of k-means.

In the second part of the chapter we study an initialization scheme for the k-means algorithm. Our intention is not to come up with a new scheme, but to show that a scheme already in use is "good" in the sense that it tends to put initial centers in different clusters. It is important to realize that such a property does not hold for the widely used uniform random initialization.

On the technical side, most of the proofs and proof ideas in this section are novel. In very broad terms, our analysis is reminiscent to that of Dasgupta and Schulman [2007]. One reason we needed new proof techniques lie partly in the fact that we analyze one-dimensional Gaussians, whose concentration properties differ qualitatively from those of high dimensional Gaussians. We loose some of the advantages high dimensionality confers. A second major difference is that k-means behaves qualitatively differently from EM whenever more than one Gaussian is involved. While EM weights a point "belonging" to a cluster by its distance to the cluster center, to the effect that far away points have a vanishing influence on a center c_j, this is not true for k-means. A far-away point can have a significative influence on the center of mass c_j, precisely because of the leverage given by the large distance. In this sense, k-means is a more brittle algorithm than EM, is less predictible and harder to analyze. In order to deal with this problem we "eliminated" impure clusters in Section 5.2. Third, while Dasgupta and Schulman [2007] is concerned with finding the correct centers when K is known, our analysis carries over to the regime when K' is too large, which is qualitatively very different of the former.

Of course many initialization schemes have been suggested and analyzed in the literature for k-means (for examples see Ostrovsky et al., 2006, Arthur and Vassilvitskii, 2007). However, these papers analyze the *clustering cost* obtained with their initialization, not the positions of the initial centers.

Part 3

Additional material and bibliography

CHAPTER 10

Statistical background

Contents

1. Concentration Inequalities

We state here all the different concentration inequalities that we use throughout the text. We start with the celebrated Hoeffding's inequality (Hoeffding [1963]).

LEMMA 10.1 (Hoeffding's Inequality). *Let X be a real random variable with $a \leqslant X \leqslant b$. Then for any $s \in \mathbb{R}$,*

$$\log\left(\mathbb{E}\exp(sX)\right) \leqslant s\mathbb{E}X + \frac{s^2(b-a)^2}{8}.$$

The second result is a consequence of Lemma 10.1 and Markov's inequality. It concerns the concentration of a sum of martingales differences.

THEOREM 10.1 (Hoeffding-Azuma's inequality for martingales). *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$ and $X_t \in [A_t, A_t + c_t]$ where A_t is a random variable $\mathcal{F}_{t-1}$-measurable and c_t is a positive constant. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geqslant \varepsilon\Big) \leqslant \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{n} c_t^2}\right), \tag{10.1}$$

or equivalently for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{n} X_t \leqslant \sqrt{\frac{\log(\delta^{-1})}{2}\sum_{t=1}^{n} c_t^2}. \tag{10.2}$$

The next result is a refinement of the previous concentration inequality which takes into account the variance of the random variables. More precisely up to a second order term it replaces the range (squared) of the random variables by their variances.

THEOREM 10.2 (Bernstein's inequality for martingales). *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$, $|X_t| \leqslant b$ for some $b > 0$ and $\mathbb{E}(X_t^2|\mathcal{F}_{t-1}) \leqslant v$ for some $v > 0$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geqslant \varepsilon\Big) \leqslant \exp\left(-\frac{\varepsilon^2}{2nv + 2b\varepsilon/3}\right), \tag{10.3}$$

and for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{n} X_t \leqslant \sqrt{2nv \log(\delta^{-1})} + \frac{b \log(\delta^{-1})}{3}. \tag{10.4}$$

PROOF. Both inequalities come from Result (1.6) of Freedman [1975]. The first inequality then uses $(1+x)\log(1+x) - x \geqslant \frac{x^2}{2+2x/3}$, while the other uses Inequality (45) of Audibert et al. [2009]. This last inequality allows to remove the $\sqrt{2}$ factor appearing in Lemma A.8 of Cesa-Bianchi and Lugosi [2006]. □

The next concentration inequality was proved by Audibert et al. [2009]. It allows to replace the true variance by its empirical estimate in Bernstein's bound.

THEOREM 10.3 (Empirical Bernstein bound). *Let $X_1, \dots, X_n$ be i.i.d centered real random variables in $[0, b]$ for some $b > 0$. Then for any $\delta > 0$ and $s \in \{1, \dots, n\}$, with probability at least $1 - \delta$, we have*

$$\sum_{t=1}^{s} X_t \leqslant \sqrt{2nV_s \log(3\delta^{-1})} + 3\log(3\delta^{-1}),$$

where $V_s = \frac{1}{s}\sum_{t=1}^{s} \left(X_t - \frac{1}{s}\sum_{\ell=1}^{s} X_\ell\right)^2$.

The above concentration inequalities were concerned with sums of random variables. There exists a vast litterature on the extension of these results to more general functionals of a sequence of random variables, see *e.g.*, Massart [2006]. We cite here a basic result in this direction which shall be enough for our purposes, namely McDiarmid's inequality (McDiarmid [1989]).

THEOREM 10.4 (McDiarmid's inequality). *Let $X_1, \dots, X_n$ be independent random variables taking values in a set A. Let $g : A^n \to \mathbb{R}$ be measurable and for any $1 \leqslant t \leqslant n$ assume that there exists a positive constant c_t such that*

$$\sup_{x_1,\dots,x_n,x' \in A} g(x_1, \dots, x_n) - g(x_1, \dots, x_{t-1}, x', x_{t+1}, \dots, x_n) \leqslant c_t.$$

Then

$$\mathbb{P}\left(|g(X_1, \dots, X_n) - \mathbb{E}g(X_1, \dots, X_n)| \geqslant \varepsilon\right) \leqslant 2\exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{n} c_t^2}\right).$$

2. Information Theory

We recall here some basic definitions and useful lemmas from Information Theory, see *e.g.*, Cover and Thomas [1991]. If $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions defined over the same sigma field and such that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ then we define the Kullback-Leibler divergence as:

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}.$$

In the case of Bernoulli distributions with parameters p and q in $(0, 1)$ we make a slight abuse of notations and note:

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

Here are three useful lemmas to compute bounds on the Kullback-Leibler divergence in different cases.

LEMMA 10.2 (Pinsker's inequality). *For any measurable set A,*

$$|\mathbb{P}(A) - \mathbb{Q}(A)| \leqslant \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}, \mathbb{Q})}.$$

LEMMA 10.3. *For any $p, q \in [0,1]$,*

$$2(p-q)^2 \leqslant \mathrm{KL}(p,q) \leqslant \frac{(p-q)^2}{q(1-q)}.$$

PROOF. The left hand side inequality is simply Lemma 10.2 for Bernoulli's random variables. The right hand side on the other hand comes from $\log x \leqslant x - 1$ and the following computations:

$$\begin{aligned}
\mathrm{KL}(p,q) &= p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\
&\leqslant p\frac{p-q}{q} + (1-p)\frac{q-p}{1-q} \\
&= \frac{(p-q)^2}{q(1-q)}.
\end{aligned}$$

□

LEMMA 10.4 (Chain Rule for Kullback-Leibler Divergence). *Assume that $\mathbb{P}$ and $\mathbb{Q}$ are defined over a finite product set $A \times B$. Let $\mathbb{P}_A, \mathbb{Q}_A$ (respectively $\mathbb{P}_B, \mathbb{Q}_B$) be the marginal distributions over A (respectively B). Then:*

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \mathrm{KL}(\mathbb{P}_B, \mathbb{Q}_B) + \int_B \mathrm{KL}(\mathbb{P}(\cdot|A \times \{b\}), \mathbb{Q}(\cdot|A \times \{b\}))d\mathbb{P}_B(b).$$

3. Probability Theory Lemma

LEMMA 10.5 (A Maximal Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of real random variables with positive mean and satisfying almost surely*

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{t=1}^{n} X_t = \mu. \tag{10.5}$$

Then we have almost surely:

$$\lim_{n\to+\infty} \frac{1}{n}\max_{1\leqslant s\leqslant n}\sum_{t=1}^{s} X_t = \mu. \tag{10.6}$$

PROOF. Let $S_n = \sum_{t=1}^{n} X_t$ and $M_n = \max_{1\leqslant i\leqslant n} S_i$. We need to prove that $\lim_{n\to+\infty} \frac{M_n}{n} = \mu$. First of all we clearly have almost surely:

$$\liminf_{n\to+\infty} \frac{M_n}{n} \geqslant \liminf_{n\to+\infty} \frac{S_n}{n} = \mu.$$

Now we need to upper bound the $\limsup$. Let $\varphi : \mathbb{N} \to \mathbb{N}$ be an increasing function such that $\varphi(n)$ is the largest integer smaller than n satisfying $M_n = S_{\varphi(n)}$. Thus

$$\frac{M_n}{n} \leqslant \frac{S_{\varphi(n)}}{\varphi(n)}.$$

If $\varphi(n) \to \infty$ then one can conclude from (10.5) that

$$\limsup_{n\to+\infty} \frac{S_{\varphi(n)}}{\varphi(n)} \leqslant \mu.$$

On the other hand if $\varphi(n) \leqslant N\ \forall n$ then for any $T > 0$ we have $\sum_{t=N+1}^{T} X_t < 0$ and this event has probability zero since $\mathbb{P}(X_t < 0) < 1$ (otherwise μ would not be positive). □

Bibliography

Y. Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, University of Alberta, 2009.

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 263–274. Omnipress, 2008.

R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Mathematics*, 27:1054–1078, 1995a.

R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995b.

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *ALT*, volume 4264 of *Lecture Notes in Computer Science*, pages 229–243. Springer, 2006.

A. Antos, V. Grover, and Cs. Szepesvári. Active learning in multi-armed bandits. In *Proc. of the 19th International Conference on Algorithmic Learning Theory*, pages 329–343. Springer-Verlag, 2008.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA*, 2007.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *22nd annual conference on learning theory*, 2009a.

J.-Y. Audibert and S. Bubeck. Minimax policies for bandits games, 2009b.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *23rd annual conference on learning theory*, 2010.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, 1995.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. *20th Conference on Learning Theory*, pages 454–468, 2007.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 89–96, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

B. Awerbuch and R.D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53. ACM, 2004.

D. J. Batstone, J. Keller, I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist, and V. A. Vavilin. Anaerobic digestion model no. 1 (adm1). *IWA Publishing*, 13, 2002.

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66:243 – 257, 2007.

S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT*, 2008.

S. Ben-David, U. von Luxburg, and D. Pál. A sober look on clustering stability. In *COLT*, 2006.

S. Ben-David, D. Pál, and H.-U. Simon. Stability of k -means clustering. In *COLT*, 2007.

S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning. In *22nd annual conference on learning theory*, 2009.

D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp. Bandit problems with infinitely many arms. *Annals of Statistics*, (25):2103–2116, 1997.

P. Billingsley. *Convergence of Probability Measures*. Wiley and Sons, 1968.

L. Bottou and Y. Bengio. Convergence properties of the k-means algorithm. In *NIPS*, 1995.

S. Bubeck and R. Munos. Open loop optimistic planning. In *23rd annual conference on learning theory*, 2010.

S. Bubeck and U. von Luxburg. Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *JMLR*, 10:657 – 698, 2009.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. of the 20th International Conference on Algorithmic Learning Theory*, 2009a.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely–armed and continuously–armed bandits, 2009b.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvari. Online optimization in $\mathcal{X}$-armed bandits. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 22*, pages 201–208, 2009c.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvari. $\mathcal{X}$-armed bandits, 2009d.

J. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical report, University of Bonn, 1998.

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. In *22nd annual conference on learning theory*, 2009.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE: Transactions on Information Theory*, 51:2152–2162, 2005.

D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal. Mortal multi-armed bandits. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 22*, pages 273–280. MIT Press, 2009.

Hyeong Soo Chang, Michael C. Fu, Jiaqiao Hu, and Steven I. Marcus. *Simulation-based Algorithms for Markov Decision Processes*. Springer, London, 2007.

GMJ Chaslot, M.H.M. Winands, H. Herik, J. Uiterwijk, and B. Bouzy. Progressive strategies for Monte-Carlo tree search. *New Mathematics and Natural Computation*, 4(3):343–357, 2008.

E. Cope. Regret and convergence bounds for immediate-reward reinforcement learning with continuous action spaces. Preprint, 2004.

P.-A. Coquelin and R. Munos. Bandit algorithms for tree search. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

A. Czumaj and C. Sohler. Sublinear-time approximation algorithms for clustering via random sampling. *Random Struct. Algorithms*, 30(1-2):226–256, 2007.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 355–366. Omnipress, 2008.

S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learnig Research*, 8:203–226, 2007.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

J. L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 255–270, 2002.

M.A.F. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3): 381–396, 2002.

S. Filippi, O. Cappé, and A. Garivier. Regret bounds for opportunistic channel access. *Available on Arxiv*, 2009.

H. Finnsson and Y. Bjornsson. Simulation-based approach to general game playing. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 259–264, 2008.

D. Foster and R. Vohra. Calibrated learning and corellated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

C. Fraley and A. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *Comput. J*, 41(8):578–588, 1998.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.

J. Fritz. Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Inf. Th.*, 21(5):552 – 557, 1975.

M. Garey, D. Johnson, and H. Witsenhausen. The complexity of the generalized Lloyd - max problem (corresp.). *IEEE Trans. Inf. Theory*, 28(2):255–256, 1982.

A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *ArXiv e-prints*, 2008.

S. Gelly and D. Silver. Achieving master level play in 9×9 computer go. In *Proceedings of AAAI*, pages 1537–1540, 2008.

S. Gelly and D. Silver. Combining online and offline knowledge in UCT. In *Proceedings of the 24th international conference on Machine learning*, pages 273–280. ACM New York, NY, USA, 2007.

S. Gelly, Y. Wang, R. Munos, and O. Teytaud. Modification of UCT with patterns in Monte-Carlo go. Technical Report RR-6062, INRIA, 2006.

J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley-Interscience series in systems and optimization. Wiley, Chichester, NY, 1989.

P. Grünwald. *The minimum description length principle*. MIT Press, Cambridge, MA, 2007.

S. Guattery and G. Miller. On the quality of spectral separators. *SIAM Journal of Matrix Anal. Appl.*, 19(3):701 – 719, 1998.

A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403, 2007.

J. Hartigan. Consistency of single linkage for high-density clusters. *JASA*, 76(374):388 – 394, 1981.

J. Hartigan. Statistical theory in clustering. *Journal of classification*, 2:63 – 76, 1985.

D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, May 1985.

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

J.-F. Hren and R. Munos. Optimistic planning for deterministic systems. In *European Workshop on Reinforcement Learning*. 2008.

D. Hsu, W.S. Lee, and N. Rong. What makes some POMDP problems easy to approximate? In *Neural Information Processing Systems*, 2007.

M. Inaba, N. Katoh, and H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based *k*-clustering. In *Proceedings of the 10th Annual Symposium on Computational Geometry*, pages 332–339. ACM Press, Stony Brook, USA, 1994.

P. Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the thirty-first annual ACM Symposium on Theory of Computing (STOC)*, pages 428–434. ACM Press, New York, 1999.

S. Jegelka. Statistical learning theory approaches to clustering. Master's thesis, University of Tübingen, 2007. Available at http://www.kyb.mpg.de/publication.html?user=jegelka.

S. M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, Gatsby Computationel Neuroscience Unit, University College London, 2003.

R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

M. Kearns, Y. Mansour, and A.Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markovian decision processes. In *Machine Learning*, volume 49, year = 2002,, pages 193–208.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *18th Advances in Neural Information Processing Systems*, 2004.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008a.

R. D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 343–354. Omnipress, 2008b.

L. Kocsis and Cs. Szepesvari. Bandit based Monte-carlo planning. In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.

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

T. Lange, V. Roth, M. Braun, and J. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299 – 1323, 2004.

A. Lazaric and R. Munos. Hybrid stochastic-adversarial online learning. In *22nd annual conference on learning theory*, 2009.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994.

K. Liu and Q. Zhao. A restless bandit formulation of opportunistic access: indexablity and index policy. In *Proc. of the 5th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON) Workshops*, 2008.

O. Madani, D. Lizotte, and R. Greiner. The budgeted multi-armed bandit problem. pages 643–645, 2004. Open problems session.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

O. Maron and A. W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *NIPS*, pages 59–66, 1993.

P. Massart. *Ecole d'Ete de Probabilites de Saint-Flour XXXIII - 2003*. Springer, Berlin, Heidelberg, New York, 2006.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148 – 188, 1989. Cambridge University Press.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley, New York, 2004.

H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *In Proceedings of the 17th Annual Conference on Learning Theory*, pages 109–123, 2004.

N. Mishra, D. Oblinger, and L. Pitt. Sublinear time approximate clustering. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-01)*, pages 439–447. ACM Press, New York, 2001.

V. Mnih, Cs. Szepesvári, and J.-Y. Audibert. Empirical bernstein stopping. In *ICML*, volume 307, pages 672–679, 2008.

M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

N.Srebro, G. Shakhnarovich, and S. Roweis. An investigation of computational and informational limits in gaussian mixture clustering. In *ICML*, 2006.

R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *FOCS*, 2006.

S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for taxonomies: A model-based approach. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, 2007a.

S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 721–728, New York, NY, USA, 2007b. ACM.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135 – 140, 1981.

A. Rakhlin and A. Caponnetto. Stability of k-means clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

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

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits, 2009.

M.P.D. Schadd, M.H.M. Winands, H.J. van den Herik, and H. Aldewereld. Addressing NP-complete puzzles with Monte-Carlo methods. In *Proceedings of the AISB 2008 Symposium on*

Logic and the Simulation of Interaction and Reasoning, volume 9, pages 55—61. The Society for the study of Artificial Intelligence and Simulation of Behaviour, 2008.

K. Schlag. Eleven tests needed for a recommendation. Technical Report ECO2006/2, European University Institute, 2006.

O. Shamir and N. Tishby. Cluster stability for finite samples. In *NIPS*. 2008a.

O. Shamir and N. Tishby. Model selection and stability in k-means clustering. In *COLT*, 2008b.

O. Shamir and N. Tishby. On the reliability of clustering stability in the large sample regime. In *NIPS*. 2008c.

A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 343–354. Omnipress, 2008.

D. Spielman and S. Teng. Spectral partitioning works: planar graphs and finite element meshes. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 96 – 105. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.

G. Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris-Sud, Orsay, France, May 2005.

M. J. Streeter and S. F. Smith. A simple distribution-free approach to the max k-armed bandit problem. *Principles and Practice of Constraint Programming (CP)*, pages 560–574, 2006a.

M. J. Streeter and S. F. Smith. An asymptotically optimal algorithm for the max k-armed bandit problem. In *AAAI*, 2006b.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395 – 416, 2007.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL workshop on Statistics and Optimization of Clustering, London*, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555 – 586, 2008.

U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 21*. MIT Press, Cambridge, MA, 2008.

D. Wagner and F. Wagner. Between min cut and graph bisection. In *Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, pages 744 – 750, London, 1993. Springer.

C.-C. Wang, S.R. Kulkarni, and H.V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

Y. Wang, J.Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1729–1736. 2009.

P. Whittle. Activity allocation in a changing world. *Journal of Applied Probability*, 27A, 1988.

M. Wong and T. Lane. A kth nearest neighbor clustering procedure. *J.R. Statist.Soc B*, 45(3): 362 – 368, 1983.

C. Yu, J. Chuang, B. Gerkey, G. Gordon, and A.Y. Ng. Open loop plans in POMDPs. Technical report, Stanford University CS Dept., 2005.

Z. Zhang, B. Dai, and A. Tung. Estimating local optimums in EM algorithm over Gaussian mixture model. In *ICML*, 2008.

Q. Zhao, L. Tong, and A. Swami. Decentralized cognitive mac for dynamic spectrum access. In *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, pages 224–232, 2005.

Jeux de Bandits et Fondations du Clustering

Résumé : Ce travail de thèse s'inscrit dans le domaine du *machine learning* et concerne plus particulièrement les sous-catégories de l'optimisation stochastique, du *online learning* et du *clustering*. Ces sous-domaines existent depuis plusieurs décennies mais ils ont tous reçu un éclairage différent au cours de ces dernières années. Notamment, les jeux de bandits offrent aujourd'hui un cadre commun pour l'optimisation stochastique et l'*online learning*. Ce point de vue conduit à de nombreuses extensions du jeu de base. C'est sur l'étude mathématique de ces jeux que se concentre la première partie de cette thèse. La seconde partie est quant à elle dédiée au *clustering* et plus particulièrement à deux notions importantes: la consistance asymptotique des algorithmes et la stabilité comme méthode de sélection de modèles.

Mots-clés : *online learning*, optimisation stochastique, jeux de bandits, apprentissage séquentiel, regret minimax, prédiction avec information incomplète, bandits avec infinité d'actions, regret non cumulé, exploration efficace, *clustering*, consistance, stabilité.

Bandits Games and Clustering Foundations

Abstract: This thesis takes place within the machine learning theory. In particular it focuses on three sub-domains, stochastic optimization, online learning and clustering. These subjects exist for decades, but all have been recently studied under a new perspective. For instance, bandits games now offer a unified framework for stochastic optimization and online learning. This point of view results in many new extensions of the basic game. In the first part of this thesis, we focus on the mathematical study of these extensions (as well as the classical game). On the other hand, in the second part we discuss two important theoretical concepts for clustering, namely the consistency of algorithms and the stability as a tool for model selection.

Keywords: online learning, stochastic optimization, bandits games, sequential learning, minimax regret, prediction with limited feedback, bandits with infinitely many arms, non-cumulative regret, efficient exploration, clustering, consistency, stability.
