# Déclinaisons de bandits et leurs applications

**Thèse**

**Audrey Durand**

**Doctorat en génie électrique**
Philosophiæ doctor (Ph.D.)

Québec, Canada

# Déclinaisons de bandits et leurs applications

**Thèse**

**Audrey Durand**

Sous la direction de:

Christian Gagné, directeur de recherche
Joelle Pineau, codirectrice de recherche

# Résumé

Cette thèse s'intéresse à différentes variantes du problème des bandits, une instance simplifiée d'un problème de *reinforcement learning* (RL) dont l'accent est mis sur le compromis entre l'exploration et l'exploitation. Plus spécifiquement, l'accent est mis sur trois variantes, soient les bandits contextuels, structurés et multi-objectifs. Dans la première, un agent recherche l'action optimale dépendant d'un contexte donné. Dans la seconde, un agent recherche l'action optimale dans un espace potentiellement grand et caractérisé par une métrique de similarité. Dans la dernière, un agent recherche le compromis optimal sur un front de Pareto selon une fonction d'articulation des préférences non observable directement. La thèse propose des algorithmes adaptés à chacune de ces variantes, dont les performances sont appuyées par des garanties théoriques ou des expériences empiriques. Ces variantes de bandits servent de cadre à deux applications réelles et à haut potentiel d'impact, soient l'allocation de traitements adaptative pour la découverte de stratégies de traitement du cancer personnalisées, et l'optimisation en-ligne de paramètres d'imagerie microscopique à grande résolution pour l'acquisition efficace d'images utilisables en neuroscience. La thèse apporte donc des contributions à la fois algorithmiques, théoriques et applicatives.

Une adaptation de l'algorithme *best empirical sampled average* (BESA), GP BESA, est proposée pour le problème des bandits contextuels. Son potentiel est mis en lumière par des expériences en simulation, lesquelles ont motivé le déploiement de la stratégie dans une étude sur des animaux en laboratoire. Les résultats, prometteurs, montrent que GP BESA est en mesure d'étendre la longévité de souris atteintes du cancer et ainsi augmenter significativement la quantité de données recueillies sur les sujets.

Une adaptation de l'algorithme Thompson *sampling* (TS), Kernel TS, est proposée pour le problème des bandits structurés en *reproducing kernel Hilbert space* (RKHS). Une analyse théorique permet d'obtenir des garanties de convergence sur le pseudo-regret cumulatif. Des résultats de concentration pour la régression à noyau avec régularisation variable ainsi qu'une procédure d'ajustement adaptative de la régularisation basée sur l'estimation empirique de la variance du bruit sont également introduits. Ces contributions permettent de lever l'hypothèse classique sur la connaissance a priori de la variance du bruit en régression à noyau en-ligne. Des résultats numériques illustrent le potentiel de ces outils. Des expériences empiriques illus-

trent également la performance de Kernel TS et permettent de soulever des questionnements intéressants relativement à l'optimalité des intuitions théoriques.

Une nouvelle variante de bandits multi-objectifs généralisant la littérature est proposée. Plus spécifiquement, le nouveau cadre considère que l'articulation des préférences entre les objectifs provient d'une fonction non observable, typiquement d'un utilisateur (expert), et suggère d'intégrer cet expert à la boucle d'apprentissage. Le concept des *rayons de préférence* est ensuite introduit pour évaluer la robustesse de la fonction de préférences de l'expert à des erreurs dans l'estimation de l'environnement. Une variante de l'algorithme TS, TS-MVN, est proposée et analysée. Des expériences empiriques appuient ces résultats et constituent une investigation préliminaire des questionnements relatifs à la présence d'un expert dans la boucle d'apprentissage.

La mise en commun des approches de bandits structurés et multi-objectifs permet de s'attaquer au problème d'optimisation des paramètres d'imagerie STED de manière en-ligne. Les résultats expérimentaux sur un vrai montage microscopique et avec de vrais échantillons neuronaux montrent que la technique proposée permet d'accélérer considérablement le processus de caractérisation des paramètres et facilitent l'obtention rapide d'images pertinentes pour des experts en neuroscience.

# Abstract

This thesis deals with various variants of the bandits problem, wihch corresponds to a simplified instance of a RL problem with emphasis on the exploration-exploitation trade-off. More specifically, the focus is on three variants: contextual, structured, and multi-objective bandits. In the first, an agent searches for the optimal action depending on a given context. In the second, an agent searches for the optimal action in a potentially large space characterized by a similarity metric. In the latter, an agent searches for the optimal trade-off on a Pareto front according to a non-observable preference function. The thesis introduces algorithms adapted to each of these variants, whose performances are supported by theoretical guarantees and/or empirical experiments. These bandit variants provide a framework for two real-world applications with high potential impact: 1) adaptive treatment allocation for the discovery of personalized cancer treatment strategies; and 2) online optimization of microscopic imaging parameters for the efficient acquisition of useful images. The thesis therefore offers both algorithmic, theoretical, and applicative contributions.

An adaptation of the BESA algorithm, GP BESA, is proposed for the problem of contextual bandits. Its potential is highlighted by simulation experiments, which motivated the deployment of the strategy in a wet lab experiment on real animals. Promising results show that GP BESA is able to extend the longevity of mice with cancer and thus significantly increase the amount of data collected on subjects.

An adaptation of the TS algorithm, Kernel TS, is proposed for the problem of structured bandits in RKHS. A theoretical analysis allows to obtain convergence guarantees on the cumulative pseudo-regret. Concentration results for the regression with variable regularization as well as a procedure for adaptive tuning of the regularization based on the empirical estimation of the noise variance are also introduced. These contributions make it possible to lift the typical assumption on the a priori knowledge of the noise variance in streaming kernel regression. Numerical results illustrate the potential of these tools. Empirical experiments also illustrate the performance of Kernel TS and raise interesting questions about the optimality of theoretical intuitions.

A new variant of multi-objective bandits, generalizing the literature, is also proposed. More specifically, the new framework considers that the preference articulation between the objec-

tives comes from a nonobservable function, typically a user (expert), and suggests integrating this expert into the learning loop. The concept of *preference radius* is then introduced to evaluate the robustness of the expert's preference function to errors in the estimation of the environment. A variant of the TS algorithm, TS-MVN, is introduced and analyzed. Empirical experiments support the theoreitcal results and provide a preliminary investigation of questions about the presence of an expert in the learning loop.

Put together, structured and multi-objective bandits approaches are then used to tackle the online STED imaging parameters optimization problem. Experimental results on a real microscopy setting and with real neural samples show that the proposed technique makes it possible to significantly accelerate the process of parameters characterization and facilitate the acquisition of images relevant to experts in neuroscience.

# Table des matières

# Liste des tableaux

# Liste des figures

# Acronymes

**ACT** *adaptive clinical trial*.

**AOM** *accousto-optic modulator*.

**BESA** *best empirical sampled average*.

**EI** écart interquartile.

**GP** *Gaussian process*.

**MCO** moindres carrés ordinaire.

**MDP** *Markov decision process*.

**MVN** normal multi-varié.

**PSF** *point spread function*.

**RCT** *randomized clinical trial*.

**RKHS** *reproducing kernel Hilbert space*.

**RL** *reinforcement learning*.

**TS** Thompson *sampling*.

**UCB** *upper confidence bound*.

Whatever you do in this life, it is not legendary unless your friends are there to see it.

---

Barney Stinson

# Remerciements

Je tiens à remercier mon directeur de thèse, Christian Gagné, pour m'avoir guidée à travers mon parcours doctoral. La liberté de recherche et la confiance qu'il a su me donner m'auront permis de me définir comme chercheure autonome. J'ai découvert en Christian un excellent superviseur, mais aussi un ami. Je tiens également à remercier ma co-directice, Joelle Pineau. Mon séjour dans son laboratoire aura été un point tournant dans mon parcours. Ses conseils et sa supervision, bien avant d'être officialisée, m'ont été d'une aide précieuse. Ses qualités de mentor m'ont permis de gagner une confiance en moi et m'ont amenée à devenir la chercheure que suis aujourd'hui. S'il m'avait été possible d'avoir un troisième superviseur, ce titre serait sans aucun doute revenu à Odalric-Ambrym Maillard. Sans lui, les contributions théoriques de cette thèse n'auraient jamais vu le jour de cette façon. Il m'a aidé à développer mes intuitions mathématiques et à les suivre. Je me suis ainsi découvert un intérêt pour les analyses théoriques, lequel a transformé ma façon de concevoir la recherche. Un merci particulier aussi à mes évaluateurs Michèle Sebag, François Laviolette et Mario Marchand, pour avoir pris le temps de lire cette thèse et avoir offert des commentaires instructifs qui auront contribué à en faire une de mes plus grandes fiertés.

Ma thèse n'aurait pas été la même sans ses contributions applicatives, lesquelles ont résulté de collaborations étroites avec des chercheurs d'autres domaines. Je me dois donc de remercier Charis Achilleos, qui s'est occupée assiduement d'administrer les traitements à nos souris expérimentales. Je dois également remercier Flavie Lavoie-Cardinal pour avoir effectué toutes les expérimentations sur le système d'imagerie STED. En plus d'être une collaboratrice hors-paire, Flavie est une amie de longue date dottée d'une curiosité scientifique qui a contribué à faire naître des projets auxquels j'ai pris grand plaisir à me consacrer. La présence de ces collègues et amis s'est avérée un support psychologique essentiel au bon déroulement de ma thèse. Un merci spécial à Max et Charles à cet effet.

Les souper, les soirées de jeux vidéos ainsi que les *weekends* de travail communs avec Charles et Gaby ont embellit mon parcours en général et m'ont permis de survivre dans les moments les plus difficiles. Un immense merci à Marion et la Rousse pour toutes les aventures et les petites attentions qui ont su alléger ces dernières années et leur conférer une *epicness* hors du commun. Ces amitiés ont contribué au maintient d'un équilibre entre ma vie et mes ambitions.

Merci à Zonzon d'avoir toujours été là, malgré cette distance qui nous séparaît. Il y a de ces relations qui semblent aptes à perdurer indéfiniment. Merci à Liliane pour son intérêt franc et sa curiosité ; par ses questions, elle aura fait naître chez moi des réflexions qui ont agrémenté les travaux présentés dans ma thèse. Merci à mes parents de m'avoir encouragée et d'avoir démontré un intérêt constant pour mes projets et mon parcours. Me sentir supportée par eux a toujours été essentiel pour moi. Finalement, merci à Marie-Lyliane pour sa présence constante, dans les bons moments comme dans les plus difficiles, à partager mes joies et mes peines. Nos interactions auront souvent fait germer dans ma têtes de nouvelles idées ou stratégies pour contourner les défis qui se dressaient devant moi. Elle a contribué à cette thèse probablement beaucoup plus qu'elle ne le croit.

# Avant-propos

Cette thèse a été réalisée sous la supervision conjointe de Christian Gagné (Université Laval) et Joelle Pineau (McGill University). Les expériences en laboratoire présentées au chapitre 2 ont été réalisées en collaboration avec Charis Achilleos (University of Cyprus), Demetris Iacovides (University of Cyprus), Katerina Strati (University of Cyprus) et Georgios Mitsis (McGill University). Les travaux présentés au chapitre 3 ont été réalisés en collaboration avec Odalric-Ambrym Maillard (INRIA). Finalement, les expériences empiriques présentées au chapitre 5 ont été réalisées en collaboration avec Flavie Lavoie-Cardinal (Université Laval) et Paul De Koninck (Université Laval).

Les travaux présentés dans cette thèse ont été financés par le conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), le Fonds de recherche du Québec — Nature et technologies (FRQ-NT), le Regroupement pour l'étude des environnements partagés intelligents répartis (REPARTI), MITACS et E Machine Learning Inc.

# Introduction

> Les seules vraies erreurs sont celles que nous commettons à répétition.
> Les autres sont des occasions d'apprentissage.
>
> (Dalaï Lama)

L'apprentissage par essais et erreurs est naturel pour les humains. Il a donné naissance à un champ de recherche en apprentissage automatique fonctionnant sur le même principe, le *reinforcement learning* (RL). Les approches de RL sont intéressantes d'un point de vue algorithmique et scientifique en raison de leurs résultats, mais également parce qu'elles sont intuitives pour les gens externes au domaine de l'apprentissage automatique. À la recherche du comportement *optimal*, un agent interagit avec son environnement et tente de s'améliorer en observant les résultats de ses actions. Il fait donc face à un compromis : essayer des actions méconnues pour parfaire sa connaissance de celles-ci au risque d'obtenir de mauvais résultats (exploration) et entreprendre une action potentiellement optimale (exploitation). Le point central en RL repose généralement sur la prise de décision séquentielle, avec des rétroactions possiblement retardées dans le temps et parfois difficiles à associer aux actions responsables. Ces mécaniques complexes rendent le compromis exploration-exploitation difficile à analyser.

Le problème des bandits (Robbins, 1952) constitue une instance simplifiée d'un problème de RL dont l'accent est entièrement mis sur le compromis exploration-exploitation. Historiquement, les *bandits* réfèrent aux machines à sous retrouvées dans un casino. Un joueur entrant dans un casino fait face à plusieurs machines (bandits) d'espérance de gain différente. Le problème de ce joueur consiste à décider quel bandit jouer dans le but de maximiser la somme totale des gains. La simplicité du jeu résultant permet de s'attaquer aux questions de fond et favorise l'obtention de garanties de convergence théoriques favorisant la compréhension des problèmes de RL plus complexes. On retrouve généralement deux types d'algorithmes de bandits : *upper confidence bound* (UCB) (Auer et al., 2002a), basé sur l'optimisme, et Thompson *sampling* (TS) (Thompson, 1933; Chapelle and Li, 2011), basé sur une intuition bayésienne. Plus récemment, *best empirical sampled average* (BESA) (Baransi et al., 2014) a également été introduit, montrant des garanties et des performances rivalisant avec les deux précédents.

## Déclinaisons de bandits

Dans cette thèse, nous nous intéressons à trois variantes du problème des bandits.

**Les bandits contextuels** Ils étendent le problème traditionnel à la situation dans laquelle l'espérance de gain associée aux différentes actions n'est pas une quantité statique. Elle est plutôt fonction d'un contexte observable au joueur. Le but consiste donc à déterminer quelle action entreprendre de manière à maximiser les gains en fonction du contexte. Autrement dit, le joueur cherche à apprendre une politique d'actions s'adaptant au contexte.

**Les bandits structurés** Ils considèrent la situation dans laquelle les actions font partie d'un espace doté d'une métrique de similarité. L'espérance de gain est alors représentée comme une fonction dans l'espace des actions. Il est donc possible de partager de l'information entre les observations acquises avec différentes actions. Cela permet de s'adapter au cas où le nombre d'actions est très grand, ce qui est problématique dans le problème des bandits traditionnel.

**Les bandits multi-objectifs** Ils considèrent la situation dans laquelle les gains à maximiser ne sont pas scalaires ; ils représentent plutôt des quantités à multiples facettes correspondant à différents objectifs à optimiser. Typiquement, ces objectifs sont en conflit, c'est-à-dire qu'il n'est pas possible de tous les maximiser/minimiser simultanément. Ce problème n'est donc pas caractérisé par solution unique, mais par un ensemble Pareto-optimal des compromis entre ces différents objectifs. Cependant, une fonction quelconque d'articulation des préférences entre les divers objectifs est caractérisée par son compromis optimal, c'est-à-dire celui qui l'optimise.

## Motivations applicatives

Depuis leur première utilisation pour la caractérisation de traitements via l'*adaptive clinical trial* (ACT) (Thompson, 1933), les problèmes de bandits ont évolué sous différentes variantes leur permettant de couvrir différentes applications. Parmi celles-ci, on retrouve la recommandation de contenu (Katariya et al., 2016), l'optimisation des hyperparamètres (Snoek et al., 2012) et l'optimisation de communication sans-fil (Maghsudi and Hossain, 2016). Les algorithmes proposés dans cette thèse seront illustrés dans les deux applications suivantes.

### Allocation de traitements adaptative

Les stratégies de traitement personnalisées donnent espoir de transformer la manière dont la médecine est pratiquée, permettant de spécialiser les stratégies de traitement aux caractéristiques individuelles d'un patient. Cette approche semble particulièrement prometteuse pour lutter contre les maladies chroniques et mortelles, car elle ouvre la porte à la prescription d'une séquence de traitements adaptés en temps réel à l'évolution de la maladie et à la

réponse du patient aux traitements précédents. Plusieurs travaux récents ont étudié l'utilisation du RL pour découvrir et optimiser automatiquement de telles stratégies de traitement séquentiel (Ernst et al., 2006; Zhao et al., 2009; Panuccio et al., 2013; Bothe et al., 2013; Escandell-Montero et al., 2014). Des résultats préliminaires indiquent que les stratégies de traitement adaptées obtenues par RL peuvent donner de meilleurs résultats que les stratégies traditionnelles non adaptatives, mais de nombreux défis demeurent avant que l'approche soit largement transférable à la pratique clinique. Un de ces défis réside dans l'acquisition d'une quantité de données nécessaire à l'apprentissage de telles politiques personnalisées. L'approche typique consiste à effectuer un *randomized clinical trial* (RCT), dans lequel les traitements sont assignés aléatoirement à des sujets animaux, puis à utiliser ses données pour élaborer une stratégie de traitements personnalisée. Cependant, comme les traitements moins efficaces conduisent à un décès plus rapide des sujets, le processus de cueillette de données peu s'avérer long et coûteux. Une stratégie alternative consiste à utiliser une procédure d'ACT pour optimiser l'efficacité de la collecter de données. Plus précisément, l'idée consiste à assigner *plus fréquemment* des traitements qui ont *plus de chance* d'être optimaux étant donné le contexte (de la maladie, du sujet, de l'environnement).

### Optimisation en-ligne de paramètres d'imagerie à super-résolution

Dans un autre ordre d'idées, la microscopie à super-résolution est une technique permettant d'observer des structures sous-cellulaires à l'échelle nanométrique (Huang et al., 2010). Des méthodes diverses ont été développées au cours de la dernière décennie pour caractériser ces structures dans les cellules vivantes. La microscopie STED (Hell and Wichmann, 1994; Willig et al., 2006) utilise des marqueurs fluorescents pour mettre en évidence une structure d'intérêt. L'acquisition d'une image STED requiert une paramétrisation importante et extrêmement difficile pour obtenir une image d'une qualité intéressante pour les chercheurs en neuroscience et en biologie. De plus, les paramètres d'imagerie *optimaux* ne sont pas constants dans le temps : ils dépendent de l'échantillon en cours, des marqueurs utilisés, ainsi que des aléas biologiques du moment. Ainsi, étant donné une nouvelle situation d'imagerie, les paramètres doivent être optimisés de manière en-ligne dans cette situation. Cela ressemble beaucoup au problème d'optimisation en-ligne des hyperparamètres (Snoek et al., 2012), dans lequel on recherche le maximum (ou le minimum) d'une fonction en s'appuyant sur les observations associées à diverses paramétrisations de la fonction. Typiquement, les observations sont bruitées, d'un bruit a priori inconnu.

Pour l'imagerie des cellules vivantes, plusieurs images doivent être prises consécutivement afin d'étudier les processus cellulaires et les changements structurels. L'imagerie d'un échantillon à plusieurs reprises avec la microscopie STED entraîne une dégradation du marqueur fluorescent et donc une perte significative du signal de fluorescence, un phénomène connu sous le nom de photoblanchiment (Staudt et al., 2011). Afin de caractériser précisément une structure

ou un processus, le contraste entre la structure et l'arrière-plan doit être suffisamment élevé pour distinguer les détails et doit donc être maximisé. Cela peut être réalisé en augmentant le temps d'imagerie ou l'intensité du laser, ce qui entraîne une augmentation du photoblanchiment et donc une perte de signal. Il peut alors être nécessaire d'optimiser les paramètres, tout en tenant compte, non seulement de la qualité des images obtenues, mais également d'autres objectifs (tel le photoblanchiment ou de temps d'imagerie), qui sont contradictoires. Ce type de problème n'est généralement pas caractérisé par une unique solution optimale, mais plutôt par un ensemble de compromis. Un utilisateur expert donné possède sa propre fonction de préférence lui permettant de choisir son compromis *préféré* correspondant à des images utilisables a posteriori.

## Questions de recherche

**Comment aborder les bandits contextuels en *Gaussian process* (GP) avec observations à retardement ?** Sous l'hypothèse que la fonction d'espérance de gain dans l'espace des contextes est issue d'un GP, des adaptations d'UCB (Krause and Ong, 2011; Valko et al., 2013) ont été proposées pour les bandits contextuels. Cependant, les stratégies UCB sont, par définition, déterministes. Il est connu (Chapelle and Li, 2011) que les stratégies stochastiques sont plus robustes à l'obtention d'observations à retardement. Une approche stochastique pour cette variante de bandits contextuels permettrait de répondre à cette question.

**Comment aborder les bandits structurés en *reproducing kernel Hilbert space* (RKHS) de bruit inconnu ?** Sous l'hypothèse que la fonction d'espérance de gain dans l'espace des actions appartient à un RKHS, des adaptations d'UCB (Srinivas et al., 2010; Valko et al., 2013) reposant sur la régression à noyau ont été proposées pour les bandits structurés. En termes algorithmiques, la régression à noyau typique implique un paramètre de régularisation qui explique à la fois la complexité de la fonction cible inconnue et la variance du bruit. Pour que les garanties théoriques (basées sur les inégalités de concentration de la régression) soient maintenues, ces approches requièrent que le paramètre de régularisation soit une quantité fixe. Malheureusement, cela implique que la variance du bruit soit connue, ce qui n'est pas réaliste en pratique. Des résultats de concentration sur la régression à noyau avec régularisation adaptative basée sur l'estimation empirique de la variance du bruit plutôt que sur sa connaissance a priori permettraient de resserrer l'écart entre la théorie et la pratique.

**Comment aborder les bandits multi-objectifs avec fonction d'articulation des préférences non observable ?** Nous abordons le problème des bandits multi-objectifs sous l'hypothèse que les préférences entre les divers objectifs sont articulées par une fonction non observable *directement*, mais dont l'option préférée peut être déterminée par un utilisateur expert parmi un ensemble d'options données. Plus spécifiquement, nous sommes intéressées

par une méthode permettant de produire les options présentées à l'expert de manière à obtenir des garanties théoriques. Cela requiert, entre autres, de caractériser la robustesse de la fonction d'articulation des préférences aux options présentées à l'expert.

## Plan

Le chapitre 1 présente une mise en contexte sur les problèmes de bandits (traditionnels) et sur les approches de régression à noyau. Les premiers servent de base à toutes les variantes présentées aux chapitres subséquents. Le chapitre 2 aborde la variante des bandits contextuels et introduit une adaptation de l'algorithme BESA pour cette situation. Le nouvel algorithme résultant est appliqué au problème d'ACT pour optimiser la cueillette de données sur de vrais animaux en laboratoires. Les chapitres 3 et 4 traitent respectivement des variantes de bandits structurés et multi-objectifs et introduisent les adaptations respectives de l'algorithme TS. Finalement, le chapitre 5 présente le pipeline complet permettant d'optimiser en ligne les paramètres d'imagerie STED, de manière à faciliter l'obtention d'images de qualité en combinant les outils de bandits structurés et multi-objectifs.

# Chapitre 1

# Notions de base

Ce chapitre introduit deux concepts fondamentaux et récurrents tout au long de la thèse, soient le problème d'optimisation des bandits ainsi que la technique de régression à noyau. Dans le but de faciliter la lecture, une revue de la littérature additionnelle spécifique aux différentes variantes de bandits est présentée indépendamment dans chacun des chapitres suivants. Nous invitons les lecteurs intéressés par un survol plus approfondi des bandits à se référer à Bubeck and Cesa-Bianchi (2012). De même, nous invitons les lecteurs intéressés à une couverture plus détaillée des approches de régression à noyau à consulter Maillard (2016) et Rasmussen and Williams (2006).

## 1.1 Les bandits

Introduit par Robbins (1952), le problème des bandits [1] est une instance simplifiée d'un problème de prise de décision dont l'emphase est mise sur le compromis exploration-exploitation. Dans ce problème d'optimisation, un agent dispose d'un ensemble d'actions disponibles pour interagir avec un environnement. Chacune de ces actions est caractérisée par une espérance de gain quelconque. Lorsqu'il entreprend une action, l'agent observe une *récompense*, un gain. Le but de l'agent est de maximiser ses gains, ce qui se traduit par une maximisation des récompenses. Idéalement, l'agent aimerait effectuer à répétition l'action *optimale*, c'est-à-dire l'action qui offre le gain le plus important. Pour ce faire, il doit acquérir assez d'information pour lui permettre de découvrir cette action optimale. Cependant, cette acquisition de données est effectuée simultanément à la prise de décision dans laquelle les gains doivent être maximisés. Le compromis réside donc entre la sélection d'actions informatives visant à parfaire la connaissance de l'environnement et la sélection d'actions permettant (potentiellement) de maximiser les gains.

On regroupe les problèmes de bandits en trois groupes, selon la provenance de leurs récom-

1. Aussi connu sous le nom de *multi-armed bandit* (MAB) ou $K$*-armed bandit*.

penses (Bubeck and Cesa-Bianchi, 2012) :

**Bandits stochastiques :** chaque action est associée à une distribution des récompenses (inconnue) stationnaire, de telle sorte que les récompenses générées par une action donnée sont i.i.d. suivant cette distribution (Robbins, 1952).

**Bandits adversariaux :** préalablement à la sélection d'une action par l'agent, les récompenses sont choisies par un adversaire et associées aux différentes actions (Auer et al., 2002b).

**Bandits markoviens :** chaque action est associée à une distribution des récompenses (inconnue) qui évolue en fonction des actions qui sont effectuées. Certains travaux considèrent que la distribution des récompenses associée à une action change seulement lorsque cette action est sélectionnée. D'autres considèrent des mécaniques d'évolution des distributions plus complexes. Dans tous les cas, ces problèmes de bandits peuvent être formulés comme des *Markov decision process*s (MDPs) complètement ou partiellement observables.

Dans la suite de la thèse, il est toujours question de bandits stochastiques. Nous omettons cependant le terme « stochastiques » pour alléger le texte.

Dans sa version originale ou *traditionnelle*, le problème des bandits (stochastiques) consiste en un jeu épisodique décrit par un ensemble d'actions $\mathcal{A}$ fini et discret. Chaque action $a \in \mathcal{A}$ est associée à une distribution $\nu_a$ d'observations d'espérance (inconnue) $\mu_a$. À chaque épisode $t \in \mathbb{N}_{>0}$, un agent choisit d'effectuer une action $a_t$ et obtient une observation $y_t \sim \nu_{a_t}$. Sans perte de généralité, il est généralement supposé que les espérances $\mu_a$ sont contenues dans l'intervalle $[0, 1]$. D'autres font également l'hypothèse d'observations bornées (Auer et al., 2002a; Valko et al., 2013).

### 1.1.1 Les métriques

Soit l'action optimale $\star := \arg\max_{a \in \mathcal{A}} \mu_a$. Le but de l'agent est de maximiser les gains obtenus avec les actions entreprises. Cette performance peut être mesurée via les récompenses obtenues. À cet effet, on distingue le regret empirique, le regret cumulatif et le pseudo-regret cumulatif. L'agent vise évidemment à minimiser ces métriques.

**Définition 1** (Regret empirique)**.** *Le regret empirique d'un agent après $T$ épisodes est défini comme*

$$\widehat{R}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} [\mu_\star - y_t].$$

**Définition 2** (Regret cumulatif)**.** *Soit $Y_{a,i}$ le $i$-ième échantillon de la distribution $\nu_a$. Le regret cumulatif d'un agent après $T$ épisodes est défini comme*

$$R(T) \stackrel{\text{def}}{=} \max_{a \in \mathcal{A}} \sum_{t=1}^{T} Y_{a,t} - \sum_{t=1}^{T} y_t.$$

**Définition 3** (Pseudo-regret cumulatif)**.** *Le pseudo-regret cumulatif d'un agent après $T$ épisodes est défini comme*

$$\mathfrak{R}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} [\mu_\star - \mu_{a_t}].$$

Les bandits (stochastiques) sont parfois abordés dans un cadre *pure exploration* (Bubeck et al., 2009) caractérisé par une phase initiale d'exploration exclusive, suivie d'une phase d'exploitation exclusive. La phase explorative correspond à un nombre d'essais (épisodes) *gratuits*, lors desquels il n'est pas nécessaire de maximiser les gains. Le but consiste exclusivement à découvrir l'action optimale, laquelle sera considérée sans exploration lors de la phase subséquente. À cet effet, le regret simple est utilisé comme métrique à minimiser.

**Définition 4** (Regret simple)**.** *Le regret simple d'un agent après $T$ épisodes d'exploration est défini comme*

$$r(T) \stackrel{\text{def}}{=} \mu_\star - \mu_{a_T}.$$

Dans la suite de la thèse, le but consistera généralement à minimiser une variante du pseudo-regret cumulatif (ou son espérance) adapté à la variante de bandits considérée. Les adaptations considérées seront introduites en temps et lieu dans chaque chapitre. L'expression « regret » sera parfois utilisée pour désigner le pseudo-regret cumulatif afin d'alléger le texte. Si la notion de pseudo-regret est plus faible que la notion de regret, au sens où $\mathfrak{R}(T) \leqslant \mathbb{E}[R(T)]$ par l'inégalité de Jensen [2], des bornes sur le pseudo-regret impliquent des bornes sur le regret pour les bandits stochastiques et pour les bandits adversariaux dont les récompenses ne dépendent pas des choix antérieurs de l'agent. Dans le premier cas, l'écart entre les bornes sur $\mathfrak{R}(T)$ et $\mathbb{E}[R(T)]$ ne correspond qu'aux fluctuations aléatoires dans les observations. Dans le deuxième cas, nous avons entre autres (Audibert and Bubeck, 2010)

$$\mathbb{E}[R(T)] - \mathfrak{R}(T) \leqslant \sqrt{\frac{T \ln |A|}{2}}. \tag{1.1}$$

**Remarque 1.** *Un algorithme uniformément aléatoire entre les actions obtiendra un pseudo-regret cumulatif linéaire. Ainsi, on vise typiquement l'obtention d'un regret sous-linéaire, plus spécifiquement de l'ordre de* $\ln(T)$ *(Lai and Robbins, 1985).*

### 1.1.2 Les algorithmes

Nous introduisons maintenant les algorithmes qui serviront de base aux approches proposées dans les variantes de bandits présentées aux chapitres suivants. Un algorithme de bandits est une méthode (possiblement randomisée) pour sélectionner la prochaine action à effectuer étant

2. Pour une variable aléatoire $X$ et une fonction convexe $\varphi$, $\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)]$.

donné l'historique des actions antérieures et des observations obtenues, $\mathcal{H}_t := \{a_s, y_s\}_{s=1}^t$. Soit le nombre de fois $N_{a,t} = \sum_{s=1}^t \mathbb{I}\{a_s = a\}$ où l'action $a$ a été sélectionnée jusqu'au temps $t$ (inclusivement). La moyenne empirique des observations pour l'action $a$ après $t$ épisodes sont respectivement données par

$$\widehat{\mu}_{a,t} = \frac{\sum_{s=1}^t \mathbb{I}\{a_s = a\} y_s}{N_{a,t}}.$$

**Optimisme face à l'incertitude**

L'algorithme *upper confidence bound* (UCB) (Auer et al., 2002a) est basé sur le principe d'« optimisme face à l'incertitude » (*optimism in the face of uncertainty*). L'idée consiste à utiliser les intervalles de confiance sur l'estimateur empirique de la moyenne de manière à obtenir une borne supérieure sur l'espérance des récompenses pour chaque action. Après une phase d'initialisation (*burn-in phase*) lors de laquelle chaque action est essayée une fois, un algorithme UCB sélectionne l'action avec la borne de confiance supérieure la plus élevée. Les intervalles de confiance suivants (voir les équations 1.2 à 1.5) sont évalués à $\infty$ pour $N_{a,t-1} = 0$, ce qui pousse UCB à sélectionner au moins une fois chaque action. L'ajout d'une phase d'initialisation ne sert donc qu'à faciliter l'évaluation numérique en débutant l'apprentissage lorsque tous les intervalles de confiance sont finis ($< \infty$). Bien entendu, la convergence de tels algorithmes est intimement liée à la qualité des intervalles de confiance.

Par exemple, la version la plus élémentaire UCB1 (Auer et al., 2002a) s'appuie sur l'inégalité de Chernoff-Hœffding (lemme 10, annexe A) et sélectionne à l'épisode $t > |\mathcal{A}|$ l'action

$$a_t = \operatorname*{arg\,max}_{a \in \mathcal{A}} \left[ \widehat{\mu}_{a,t-1} + \sqrt{\frac{2 \ln t}{N_{a,t-1}}} \right]. \tag{1.2}$$

Sous l'hypothèse d'observations contenues dans l'intervalle $[0, 1]$, l'intervalle de confiance

$$\mathbb{P}\left[ |\widehat{\mu}_{a,t-1} - \mu_a| \geqslant \sqrt{\frac{2 \ln t}{N_{a,t-1}}} \right] \leqslant 2t^{-4}$$

permet d'obtenir des bornes de convergence sur l'espérance du pseudo-regret.

La variance des observations a ensuite été incorporée dans la borne de confiance. L'algorithme UCB1-Tuned (Auer et al., 2002a) sélectionne à l'épisode $t > |\mathcal{A}|$ l'action

$$a_t = \operatorname*{arg\,max}_{a \in \mathcal{A}} \left[ \widehat{\mu}_{a,t-1} + \sqrt{\frac{\ln t}{N_{a,t-1}} \min\{1/4, V_{a,t-1}\}} \right], \quad \text{où} \tag{1.3}$$

$$V_{a,t} = \left( \frac{1}{N_{a,t}} \sum_{s=1}^t \mathbb{I}\{a_s = a\} y_s^2 \right) - \widehat{\mu}_{a,t}^2 + \sqrt{\frac{2 \ln t}{N_{a,t}}}$$

est une borne de confiance supérieure sur la variance des observations associées à l'action $a$ et le facteur $1/4$ correspond à une borne supérieure sur la variance d'une variable aléatoire suivant une distribution Bernoulli. Notons qu'UCB1-Tuned ne possède pas de bornes supérieures sur

le regret prouvées. Similairement, l'algorithme UCB-V (Audibert et al., 2009) sélectionne à l'épisode $t > |\mathcal{A}|$ l'action

$$a_t = \arg\max_{a \in \mathcal{A}} \left[ \widehat{\mu}_{a,t-1} + \sqrt{\frac{2\zeta \widehat{\sigma}^2_{a,t-1} \ln t}{N_{a,t-1}}} + c \frac{3\zeta \ln t}{N_{a,t-1}} \right], \quad \text{où} \tag{1.4}$$
$$\widehat{\sigma}^2_{a,t} = \frac{\sum_{s=1}^{t} \mathbb{I}\{a_s = a\}(y_s - \widehat{\mu}_{a,t})^2}{N_{a,t}}$$

est l'estimateur (biaisé) de la variance des observations et $c, \zeta \geqslant 0$. Des bornes supérieures sur l'espérance du pseudo-regret sont fournies, par exemple pour $c = 1$ et $\zeta = 1.2$.

Finalement, l'algorithme UCB1-Normal (Auer et al., 2002a) étend UCB1 au cas où observations sont échantillonnées d'une distribution normale et KL-UCB (Garivier, 2011; Maillard et al., 2011) supporte n'importe quelle famille de distribution paramétrique. Plus spécifiquement, KL-UCB s'appuie sur la divergence Kullback-Leibler pour construire l'intervalle de confiance et sélectionne à l'épisode $t > |\mathcal{A}|$ l'action

$$a_t = \arg\max_{a \in \mathcal{A}} \sup \left\{ q \in \Theta : \mathrm{KL}_\nu(\widehat{\mu}_{a,t-1}, q) \leqslant \frac{f(t-1)}{N_{a,t-1}} \right\}, \tag{1.5}$$

où $\Theta$ est l'espace tel que $\mu_a \in \Theta$ et $f$ est une fonction non décroissante, typiquement de l'ordre $f(t) \approx \ln t$ (Cappé et al., 2013). Par exemple, pour des récompenses comprises dans l'intervalle $[0, 1]$, $\Theta = [0, 1]$ et $f(t) = \ln t + C \ln(\ln t)$ pour $C \geqslant 0$. Pour des distributions d'observations Bernoulli, pour $p, q \in [0, 1]^2$, la divergence Kullback-Leibler est donnée par

$$\mathrm{KL}_{\mathcal{B}}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q}.$$

**Intuition bayésienne**

L'algorithme Thompson *sampling* (TS), initialement introduit par Thompson (1933), est une approche bayésienne aussi connue sous le nom de *randomized probability matching* (Scott, 2010). L'idée générale consiste à sélectionner la prochaine action selon sa probabilité a posteriori d'être optimale. En assumant une distribution a priori simple sur les paramètres de distribution des observations pour chaque action, la distribution postérieure étant donné les observations précédentes peut être calculée de manière bayésienne.

Soit $Y_{a,i}$ le $i$-ième échantillon associé à l'action $a$. Suivant le théorème de Bayes, la distribution postérieure sur l'espérance $\mu_a$ étant donné les observations $Y_{a,1}, \ldots, Y_{a,N}$ est donnée par

$$\mathbb{P}[\mu_a | Y_{a,1}, \ldots, Y_{a,N}] = \frac{\mathbb{P}[Y_{a,1}, \ldots, Y_{a,N} | \mu_a] \mathbb{P}[\mu_a]}{\int \mathbb{P}[Y_{a,1}, \ldots, Y_{a,N} | \mu'] \mathbb{P}[\mu'] \, \mathrm{d}\mu'}.$$

La vraisemblance des observations, $\mathbb{P}[Y_{a,1}, \ldots, Y_{a,N} | \mu_a]$, est généralement déterminée par une hypothèse sur le processus de génération des observations.

Tableau 1.1 – Distributions a priori conjuguées pour certaines fonctions de vraisemblance.

| Distribution | Paramètre | Postérieure après $N$ observations |
|---|---|---|
| Bernoulli | $p$ | $\mathrm{Beta}(\alpha_0 + \sum_{i=1}^{N} Y_{a,i}, \beta_0 + \sum_{i=1}^{N}(1 - Y_{a,i}))$ [3] |
| Normale | $\mu$ | $\mathcal{N}\left(\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{N} Y_{a,i}}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1}\right)$ [4] |
| Poisson | $\lambda$ | $\Gamma(\alpha_0 + \sum_{i=1}^{N} Y_{a,i}, \beta_0 + N)$ |

Une distribution a priori de la même forme algébrique que la postérieure est dite une *a priori conjuguée*. L'utilisation d'une distribution a priori conjuguée permet d'exprimer la distribution postérieure en forme analytique (fermée). Celle-ci permet d'éviter l'évaluation de l'intégrale, qui peut être plus ou moins difficile à évaluer. Toutes les distributions membres de la famille exponentielle possèdent des distributions a priori conjuguées. Le tableau 1.1 en donne quelques exemples pour certaines fonctions de vraisemblance.

Typiquement, TS maintient, pour chaque action $a \in \mathcal{A}$, une distribution postérieure $\pi_{a,t}$ sur l'espérance des observations pour cette action. À l'épisode $t$, TS échantillonne une valeur $\theta_{a,t} \sim \pi_{a,t-1}$ pour chaque action $a \in \mathcal{A}$ et sélectionne

$$a_t = \arg\max_{a \in \mathcal{A}} \theta_{a,t}. \tag{1.6}$$

Cette approche a dû attendre plusieurs années avant de gagner en popularité, faute de garanties théoriques sur la convergence du regret (Chapelle and Li, 2011). Dans les dernières années, TS a reçu beaucoup d'attention et plusieurs analyses du regret ont été proposées (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013; Russo and Van Roy, 2014, 2016). Les algorithmes 1 et 2 montrent le déroulement de TS d'a priori Bernoulli (avec $\alpha_0 = \beta_0 = 1$) et TS d'a priori normal (avec $\mu_0 = 0$, $\sigma_0 = 1$ et $\sigma^2 = 1$). Ainsi, selon l'intuition bayésienne, une action $a$ est sélectionnée selon sa probabilité d'être optimale, c'est-à-dire $\mathbb{P}[a = a_t] = \mathbb{P}[a = \star]$.

### Sous-échantillonnage équitable

Plus récemment, Baransi et al. (2014) ont introduit un nouvel algorithme de bandits : *best empirical sampled average* (BESA). L'idée se base sur la nécessité d'effectuer une comparaison équitable des moyennes empiriques des différentes actions. Soient les deux actions $a, b \in \mathcal{A}$ qui ont été essayées respectivement $N_a$ et $N_b$ fois, avec $N_a > N_b$. Comparer leurs moyennes empiriques serait *injuste* parce que, l'action $a$ disposant de plus d'échantillons que l'action $b$, les intervalles de confiance de leurs estimateurs empiriques de la moyenne ne sont pas comparables.

3. Des valeurs typiques d'a priori sont $\alpha_0 = \beta_0 = 1$, ce qui correspond à une distribution initiale uniforme dans l'intervalle $[0, 1]$ (Chapelle and Li, 2011).

4. Les paramètres a priori $\mu_0$ et $\sigma_0$ correspondent typiquement à la moyenne et à l'écart type des données. D'autres valeurs plus arbitraires telles $\mu_0 = 0$ et $\sigma_0 = 1$ peuvent également être considérées. De plus, il faut noter qu'il est assumé que la variance $\sigma^2$ est connue. Une pratique consiste à fixer $\sigma^2 = 1$ lorsqu'elle est inconnue (Agrawal and Goyal, 2013).

**Algorithme 1** TS-Bernoulli (Agrawal and Goyal, 2013)

**for all** épisode $t \geqslant 1$ **do**
  **for all** action $a \in \mathcal{A}$ **do**
    échantillonner $\theta_{a,t} \sim \text{Beta}\Big(1 + \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} y_s, 1 + N_{a,t-1} - \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} y_s\Big)$
  **end for**
  jouer $a_t = \arg\max_{a \in \mathcal{A}} \theta_{a,t}$ et observer $y_t$
**end for**

**Algorithme 2** TS-Normal (Agrawal and Goyal, 2013)

**for all** épisode $t \geqslant 1$ **do**
  **for all** action $a \in \mathcal{A}$ **do**
    échantillonner $\theta_{a,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} y_s}{N_{a,t-1}+1}, \frac{1}{N_{a,t-1}+1}\right)$
  **end for**
  jouer $a_t = \arg\max_{a \in \mathcal{A}} \theta_{a,t}$ et observer $y_t$
**end for**

**Algorithme 3** Sélection BESA pour deux actions, $\texttt{BESA}(a, b)$ (Baransi et al., 2014)

Paramètres : épisode en cours $t$, deux actions $a$ et $b$

$N = \min(N_{a,t-1}, N_{b,t-1})$
$\mathcal{I}_a \leftarrow \texttt{subsample}(N, N_{a,t-1})$ et $\mathcal{I}_b \leftarrow \texttt{subsample}(N, N_{b,t-1})$
calculer $\tilde{m}_{a,t} = \frac{\sum_{i \in \mathcal{I}_a} Y_{a,i}}{N}$ et $\tilde{m}_{b,t} = \frac{\sum_{i \in \mathcal{I}_b} Y_{b,i}}{N}$
choisir $a_t = \arg\max_{i \in \{a,b\}} \tilde{m}_{i,t}$  (briser l'égalité avec $a_t = \arg\min_{i \in \{a,b\}} N_{i,t-1}$)
**return** $a_t$

Pour compenser cette situation, BESA sous-échantillonne sans remplacement $N_b$ observations parmi les $N_a$ observations disponibles pour l'action $a$ et calcule la moyenne empirique de $a$ sur ce sous-ensemble. Il sélectionne ensuite l'action maximisant la moyenne empirique. Par cette procédure, BESA garantit une exploration proportionnelle à la probabilité de chaque action d'être l'action optimale.

Soit la fonction de sous-échantillonnage de $n$ parmi $m$, $\texttt{subsample}(n, m)$ pour $n, m \in \mathbb{N}$, telle que $\texttt{subsample}(n, m) = \{1, \ldots, m\}$ si $n \geqslant m$. Soient $Y_{a,i}$ le $i$-ème échantillon associé à l'action $a$ et $\mathcal{Y}_{a,t} := \{Y_{a,1}, \ldots, Y_{a,N_{a,t}}\}$ l'ensemble des observations antérieures associées à l'action $a$ jusqu'au temps $t$ (inclusivement). L'algorithme 3 décrit la procédure pour sélectionner l'action à effectuer au temps $t$ parmi deux actions. Soulignons qu'en cas d'égalité entre $\tilde{m}_{a,t}$ et $\tilde{m}_{b,t}$, l'algorithme opte pour l'exploration en sélectionnant l'action la moins jouée. L'algorithme 4 montre comment étendre cette procédure à plus de deux actions via une approche par tournoi. Pour éviter d'introduire un biais, Baransi et al. (2014) recommandent de mélanger aléatoirement l'ensemble $\mathcal{K}$ chaque fois que l'algorithme 4 est appelé.

**Algorithme 4** Tournoi BESA pour un ensemble d'actions, $\texttt{BESA}(\mathcal{K})$ (Baransi et al., 2014)

Paramètres : épisode en cours $t$, ensemble d'actions $\mathcal{K} = \{k_i\}_{1 \leqslant i \leqslant K}$ de taille $K$

**if** $\mathcal{K} = \{k\}$ **then**
  $a_t = k$
**else**
  $a_t = \texttt{BESA}\Big(\texttt{BESA}\Big(\{k_i\}_{1 \leqslant i \leqslant \lfloor \frac{K}{2} \rfloor}\Big), \texttt{BESA}\Big(\{k_i\}_{\lceil \frac{K}{2} \rceil \leqslant i \leqslant K}\Big)\Big)$
**end if**
**return** $a_t$

### 1.1.3 Les analyses théoriques

La littérature sur les bandits est en grande partie axée sur les contributions théoriques visant à analyser la convergence du regret. Nous présentons alors quelques techniques d'analyse du regret retrouvées dans la littérature. Plus spécifiquement, nous considérons ici le pseudo-regret cumulatif (ou son espérance), qui est utilisé dans la suite de la thèse.

Rappelons que $\mathcal{H}_t$ et $N_{a,t}$ dénotent respectivement l'historique des actions et observations ainsi que le nombre de fois où l'action $a$ a été effectuée, jusqu'au temps $t$ (inclusivement). Soit $U_{a,t}(\delta)$ une borne de confiance supérieure sur l'estimateur de l'espérance des récompenses de l'action $a$ au temps $t$, tel que

$$\mathbb{P}[\mu_a > U_{a,t} | \mathcal{H}_{t-1}] \leqslant \delta,$$

pour $\delta \in [0, 1]$.

#### Contrôle des tirages

Considérons la décomposition de l'espérance du pseudo-regret cumulatif (définition 3) :

$$\begin{aligned}
\mathbb{E}[\mathfrak{R}(T)] &= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathbb{P}[a_t = a](\mu_\star - \mu_a) \\
&= \sum_{a \in \mathcal{A}} (\mu_\star - \mu_a) \sum_{t=1}^{T} \mathbb{P}[a_t = a] \\
&= \sum_{a \in \mathcal{A}} (\mu_\star - \mu_a) \mathbb{E}[N_{a,T}].
\end{aligned}$$

Une technique classique (Auer et al., 2002a; Audibert et al., 2009; Maillard et al., 2011; Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013; Baransi et al., 2014) utilisée pour l'obtention de bornes sur ce regret consiste à contrôler l'espérance du nombre de sélections de chaque action sous-optimale. Une approche de ce type est considérée au chapitre 4.

Pour un algorithme UCB, le nombre de tirages d'une action sous-optimale dépend de la vitesse de convergence des intervalles de confiance utilisés dans la procédure. Rappelons qu'au

temps $t$, UCB sélectionne $a_t = \arg\max_{a\in\mathcal{A}} U_{a,t}$. Ainsi, pour une action $a$ sous-optimale,

$$\begin{aligned}\mathbb{E}[N_{a,T}] &\leqslant \mathbb{E}\Big[1 + \sum_{t=|\mathcal{A}|+1}^{T} \mathbb{I}\{U_{a,t} \geqslant U_{\star,t}\}\Big] \\ &\leqslant \mathbb{E}\Big[\ell + \sum_{t=|\mathcal{A}|+1}^{T} \mathbb{I}\{U_{a,t} \geqslant U_{\star,t}, N_{a,t-1} > \ell\}\Big] \\ &\leqslant \ell + \sum_{t=|\mathcal{A}|+1}^{T} \mathbb{P}[U_{a,t} \geqslant U_{\star,t}, N_{a,t-1} > \ell]. \end{aligned} \tag{1.7}$$

Rappelons que l'algorithme UCB1 (équation 1.2) considère $U_{a,t} = \widehat{\mu}_{a,t-1} + \sqrt{2\ln t/N_{a,t-1}}$. Ainsi, la situation où $U_{a,t} \geqslant U_{\star,t}$ implique au moins l'une des trois conditions suivantes :

$$\widehat{\mu}_{\star,t-1} \leqslant \mu_\star - \sqrt{2\ln t/N_{\star,t-1}} \qquad \widehat{\mu}_{a,t-1} \geqslant \mu_a + \sqrt{2\ln t/N_{a,t-1}} \qquad \mu_\star < \mu_a + 2\sqrt{2\ln t/N_{a,t-1}}.$$

Le contrôle des deux premières conditions repose alors sur les intervalles de confiance et la dernière condition est invalidée pour $\ell \geqslant 8\ln t/\Delta_a^2$. Cela montre bien que le pseudo-regret cumulatif est directement lié à la convergence des intervalles de confiance.

Pour un algorithme TS, le contrôle du nombre de tirages d'une action sous-optimale s'appuie sur des inégalités de concentration. Rappelons qu'au temps $t$, TS sélectionne $a_t = \arg\max_{a\in\mathcal{A}} \theta_{a,t}$, avec $\theta_{a,t} \sim \pi_{a,t-1}$, où $\pi_{a,t-1}$ est la distribution postérieure de l'espérance des observations pour l'action $a$ conditionnée sur les données obtenues jusqu'à l'épisode $t-1$ (inclusivement). Pour une action $a \in \mathcal{A}$ quelconque, soient $E^{\mu}_{a,t}$ et $E^{\theta}_{a,t}$ les événements où l'espérance postérieure est proche de la vraie moyenne, $|\mathbb{E}[\pi_{a,t-1}] - \mu_a| < C^{\mu}_{a,t}$, et où l'échantillon est proche de l'espérance postérieure, $|\theta_{a,t} - \mathbb{E}[\pi_{a,t-1}]| < C^{\theta}_{a,t}$, pour des seuils $C^{\mu}_{a,t}$ et $C^{\theta}_{a,t}$ donnés. Pour faire le lien avec UCB, nous avons alors $\mathbb{E}[\pi_{a,t-1}] + C_{a,t} = U_{a,t}$. Pour une action sous-optimale $a$, on peut décomposer

$$\sum_{t=1}^{T} \mathbb{P}[a_t = a] = \sum_{t=1}^{T} \mathbb{P}[a_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t}] + \sum_{t=1}^{T} \mathbb{P}[a_t = a, E^{\mu}_{a,t}, \overline{E^{\theta}_{a,t}}] + \sum_{t=1}^{T} \mathbb{P}[a_t = a, \overline{E^{\mu}_{a,t}}] \tag{1.8}$$

et contrôler indépendamment chaque partie. Le premier terme dépend d'un mauvais échantillonnage de l'action optimale. Le second est contrôlé par la probabilité d'un mauvais échantillonnage de l'action $a$. Le dernier est contrôlé par la concentration de l'estimateur $\mathbb{E}[\pi_{a,t-1}]$. Si les bornes obtenues en théorie dépendent des outils mathématiques permettant de caractériser la concentration de la distribution postérieure ainsi que la concentration de l'échantillonnage, la convergence de ces termes, dans les faits, dépend plutôt de la meilleure concentration *possible* de la distribution et de l'échantillonnage.

Finalement, pour un algorithme BESA, le contrôle du nombre de tirages d'une action sous-optimale $a$ s'appuie sur le nombre de tirages consécutifs de $a$. Considérons le cas à deux actions, $\mathcal{A} = \{a, \star\}$. Rappelons qu'au temps $t$, BESA sélectionne $a_t = \arg\max_{i\in\mathcal{A}} \tilde{m}_{i,t}$, où $\tilde{m}_{i,t}$ est la

moyenne empirique calculée sur le sous-échantillon associé à l'action $i$ au temps $t$. On peut alors décomposer

$$\begin{aligned}
\mathbb{E}[N_{a,T}] &= \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{\tilde{m}_{a,t} > \tilde{m}_{\star,t}\}\Big] \\
&= \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{\tilde{m}_{a,t} > \tilde{m}_{\star,t}, N_{\star,t-1} > \ell\}\mathbb{I}\{N_{\star,t-1} > \ell\} \\
&\qquad + \sum_{t=1}^{T} \mathbb{I}\{\tilde{m}_{a,t} > \tilde{m}_{\star,t}, N_{\star,t-1} \leqslant \ell\}\mathbb{I}\{N_{\star,t-1} \leqslant \ell\}\Big] \\
&\leqslant \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{\tilde{m}_{a,t} > \tilde{m}_{\star,t}, N_{\star,t-1} > \ell\} + \sum_{t=1}^{T} \mathbb{I}\{N_{\star,t-1} \leqslant \ell\}\Big] \\
&\leqslant \sum_{t=1}^{T} \mathbb{P}[\tilde{m}_{a,t} > \tilde{m}_{\star,t}, N_{\star,t-1} > \ell] + \sum_{t=1}^{T} \mathbb{P}[N_{\star,t-1} \leqslant \ell]. \qquad (1.9)
\end{aligned}$$

Il s'agit alors de montrer que le regret est contrôlé avec *suffisamment* ($\ell$) d'observations de l'action optimale (première somme) et que BESA est en mesure de choisir l'action optimale suffisamment pour atteindre cette quantité (deuxième somme). Ces deux points reposent respectivement sur la concentration de l'échantillonnage de sous-ensembles et sur le nombre maximal d'échantillons consécutifs de l'action sous-optimale. Similairement à TS, la convergence théorique de BESA dépend des outils mathématiques de concentration alors que dans les faits, la convergence de BESA dépend de la meilleure concentration *possible* de l'échantillonnage de sous-ensembles.

### Contrôle des écarts de sous-optimalité

Plutôt que de contrôler $\sum_{t=1}^{T} \mathbb{P}[a_t = a]$, ou $\mathbb{E}[N_{a,t}]$, pour chaque action $a$ sous-optimale, d'autres analyses (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016) présentent des bornes sur le pseudo-regret cumulatif (définition 3) par le contrôle de l'écart de sous-optimalité immédiat, $[\mu_\star - \mu_{a_t}]$. Une approche de ce type est considérée au chapitre 3.

Une telle analyse peut être effectuée en séparant les actions en deux groupes : celles dont l'écart de sous-optimalité est supérieur à l'intervalle de confiance, et les autres. Par exemple, rappelons que TS sélectionne $a_t = \arg\max_{a\in\mathcal{A}} \theta_{a,t}$, avec $\theta_{a,t} \sim \pi_{a,t-1}$, où $\pi_{a,t-1}$ est la distribution postérieure conditionnée sur les données obtenues jusqu'à l'épisode $t-1$ (inclusivement). Soit les seuils $C^{\mu}_{a,t}$ et $C^{\theta}_{a,t}$ tels que

$$\big|\mathbb{E}[\pi_{a,t-1}] - \mu_a\big| < C^{\mu}_{a,t} \quad \text{et} \quad \big|\theta_{a,t} - \mathbb{E}[\pi_{a,t-1}]\big| < C^{\theta}_{a,t}.$$

Par définition, aucune action $a$ (nécessairement) sous-optimale telle que $[\mu_\star - \mu_a] > C^{\mu}_{a,t} + C^{\theta}_{a,t}$, ne peut être sélectionnée si $\theta_\star \geqslant \mu_\star$. L'anti-concentration [5] de l'échantillonnage peut garantir

5. L'anti-concentration borne inférieurement la probabilité qu'un échantillon soit loin de la moyenne de sa distribution. On peut le voir comme l'opposé de la concentration, qui borne supérieurement la même probabilité.

cet événement avec une certaine probabilité. L'écart de sous-optimalité des autres actions, soit celles pour lesquelles $[\mu_\star - \mu_a] \leqslant C^\mu_{a,t} + C^\theta_{a,t}$, est borné par définition de $C^\mu_{a,t}$ et $C^\theta_{a,t}$. Chaque fois qu'une action $a$ (nécessairement sous-optimale) du premier groupe est sélectionnée, la convergence de $C^\mu_{a,t}$ et $C^\theta_{a,t}$ réduit ses chances d'être sélectionnée une prochaine fois. Chaque fois qu'une action sous-optimale $a$ du deuxième groupe est sélectionnée, ses intervalles $C^\mu_{a,t}$ et $C^\theta_{a,t}$ rétrécissent, jusqu'à faire passer l'action dans le premier groupe. Naturellement, plus l'écart de sous-optimalité est mince, plus le passage au premier groupe est retardé. Ceci dit, la quantité de pseudo-regret accumulée lors de la sélection d'une action à faible écart de sous-optimalité est également faible.

Cette analyse montre bien la capacité de TS à éliminer au fil des essais les actions sous-optimales de plus en plus près de l'optimum.

#### Contrôle du regret bayésien

Russo and Van Roy (2014) présentent une analyse basée sur le regret bayésien. Cette dernière permet d'illustrer les liens entre le regret des approches UCB et TS. Considérons un ensemble de fonctions $\mathcal{F} := \{f_\theta : \mathcal{A} \mapsto \mathbb{R} | \theta \in \Theta\}$ caractérisant les observations obtenues, tel que $\mu_a = f_\theta(a)$ pour la paramétrisation (inconnue) $\theta$ du problème. L'espérance du pseudo-regret cumulatif (définition 3) peut donc s'exprimer comme suit :

$$\mathbb{E}[\mathfrak{R}(T, \theta)] \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{E}[f_\theta(\star) - f_\theta(a_t) | \theta]. \tag{1.10}$$

Le regret bayésien (ou espérance bayésienne du pseudo-regret cumulatif dans la terminologie de la thèse) est alors donné par

$$\mathfrak{R}^{\text{bayes}}(T) \overset{\text{def}}{=} \mathbb{E}_{\theta \in \Theta}\left[\mathbb{E}[\mathfrak{R}(T, \theta)]\right] = \sum_{t=1}^{T} \mathbb{E}[f_\theta(\star) - f_\theta(a_t)], \tag{1.11}$$

où l'espérance est effectuée par rapport à la distribution a priori sur $\theta$.

Par définition d'un algorithme UCB, $a_t = \arg\max_{a \in \mathcal{A}} U_{a,t}$. Le pseudo-regret immédiat d'un UCB se décompose alors comme

$$\begin{aligned} f_\theta(\star) - f_\theta(a_t) &= f_\theta(\star) - U_{a_t,t} + U_{a_t,t} - f_\theta(a_t) \\ &\leqslant [f_\theta(\star) - U_{\star,t}] + [U_{a_t,t} - f_\theta(a_t)], \end{aligned} \tag{1.12}$$

où l'inégalité utilise le fait que $U_{a_t,t} \geqslant U_{\star,t}$ (par définition de l'algorithme). Par définition de la borne de confiance, le premier terme est négatif avec grande probabilité $1 - \delta$. Le second terme pénalise le regret pour à la lâcheté de ses intervalles de confiance. Cela montre bien que la puissance d'une approche UCB se mesure à l'étroitesse des intervalles de confiance disponibles. En considérant la situation où $f_\theta \in [0, C]$, le regret bayésien d'un UCB s'exprime

comme

$$\mathfrak{R}^{\text{bayes}}(T) \leqslant \sum_{t=1}^{T} \mathbb{E}[U_{a_t,t} - f_\theta(a_t)] + C \sum_{t=1}^{T} \mathbb{P}[f_\theta(\star) > U_{\star,t}]. \tag{1.13}$$

Par définition du TS, l'intuition bayésienne suggère que $\mathbb{P}[a = a_t] = \mathbb{P}[a = \star]$. Ainsi, puisque $U_{a,t}$ est déterministe sachant $\mathcal{H}_{t-1}$, nous avons

$$\begin{aligned}
\mathbb{E}[U_{a_t,t}|\mathcal{H}_{t-1}] &= \sum_{a \in \mathcal{A}} U_{a,t} \mathbb{P}[a = a_t|\mathcal{H}_{t-1}] \\
&= \sum_{a \in \mathcal{A}} U_{a,t} \mathbb{P}[a = \star|\mathcal{H}_{t-1}] \\
&= \mathbb{E}[U_{\star,t}|\mathcal{H}_{t-1}].
\end{aligned}$$

Le regret bayésien immédiat de TS se décompose donc comme

$$\begin{aligned}
\mathbb{E}[f_\theta(\star) - f_\theta(a_t)] &= \mathbb{E}\big[\mathbb{E}[f_\theta(\star) - f_\theta(a_t)|\mathcal{H}_{t-1}]\big] \\
&= \mathbb{E}\big[\mathbb{E}[f_\theta(\star) + U_{a_t,t} - U_{\star,t} - f_\theta(a_t)|\mathcal{H}_{t-1}]\big] \\
&= \mathbb{E}\big[\mathbb{E}[f_\theta(\star) - U_{\star,t}|\mathcal{H}_{t-1}] + \mathbb{E}[U_{a_t,t} - f_\theta(a_t)|\mathcal{H}_{t-1}]\big] \\
&= \mathbb{E}[f_\theta(\star) - U_{\star,t}] + \mathbb{E}[U_{a_t,t} - f_\theta(a_t)].
\end{aligned} \tag{1.14}$$

En considérant la situation où $f_\theta \in [0, C]$, le regret bayésien de TS s'exprime comme

$$\mathfrak{R}^{\text{bayes}}(T) \leqslant \sum_{t=1}^{T} \mathbb{E}[U_{a_t,t} - f_\theta(a_t)] + C \sum_{t=1}^{T} \mathbb{P}[f_\theta(\star) > U_{\star,t}]. \tag{1.15}$$

Si les équations 1.13 et 1.15 semblent identiques, il n'en demeure pas moins que la borne sur le regret bayésien d'UCB dépend des intervalles de confiance connus utilisés dans la politique UCB. Contrairement, le regret bayésien de TS dépend des *meilleurs* intervalles de confiance possible. Ainsi, Russo and Van Roy (2014) mettent l'emphase sur la puissance de TS, dont la performance ne dépend pas de résultats théoriques existants sur les intervalles de confiance, lesquels peuvent être difficiles à obtenir, par exemple en présence de relations complexes entre les actions. Cela conclut le survol des techniques d'analyse du pseudo-regret cumulatif.

## 1.2 La régression à noyau

Cette section présente maintenant un outil mathématique en vogue sur lequel s'appuient les approches présentées aux chapitres 2 et 3. Dans de nombreuses applications telles que l'optimisation des hyperparamètres (Snoek et al., 2012), l'apprentissage actif de préférences (Brochu et al., 2008) et l'apprentissage par renforcement (Marchant and Ramos, 2014; Wilson et al., 2014), la prise de décision s'appuie sur un modèle d'une fonction de récompenses. Les approches de régression permettent de modéliser des fonctions (possiblement non linéaires) en

partageant l'information recueillie à travers les observations, ce qui les rend très attrayantes. Elles ont d'ailleurs reçu beaucoup d'attention dans différentes variantes de problèmes de bandits, tels que les bandits contextuels et les bandits structurés, respectivement traités aux chapitres 2 et 3. Concrètement, la régression consiste à déterminer la relation entre les variables d'entrée d'une fonction et les variables en sortie de la même fonction.

### 1.2.1 Régression linéaire

Considérons pour débuter le cas simple du modèle de régression linéaire. Soit une fonction linéaire $f$ décrite par un paramètre $\theta^\top \in \mathbb{R}^d$ et définie sur un ensemble compact $\mathcal{X} \subset \mathbb{R}^d$, pour $d \in \mathbb{N}$, telle que $f(x) = \theta^\top x$ pour tout $x \in \mathcal{X}$. Soient des observations bruitées $\mathbf{y}_N = (y_1, \dots, y_N)^\top \in \mathbb{R}^N$ de la fonction $f$ obtenues aux points $\mathbf{X}_N = (x_1, \dots, x_N)$, tel que $y_i = f(x_i) + \xi_i$, de bruit $\xi_i$. L'estimateur des moindres carrés ordinaire (MCO) est une des méthodes les plus communes pour estimer $\theta$ étant donné les observations antérieures $\mathbf{X}_N$ et $\mathbf{y}_N$. Le principe consiste à considérer l'estimateur $\theta_N$ minimisant la somme des écarts au carré

$$\begin{aligned}
\theta_N &= \operatorname*{arg\,min}_{\theta' \in \mathbb{R}^{d\times 1}} \sum_{i=1}^{N} \left(y_i - x_i^\top \theta'\right)^2 \\
&= \operatorname*{arg\,min}_{\theta' \in \mathbb{R}^{d\times 1}} (\mathbf{y}_N - \mathbf{X}_N\theta')^\top (\mathbf{y}_N - \mathbf{X}_N\theta') \\
&= \operatorname*{arg\,min}_{\theta' \in \mathbb{R}^{d\times 1}} \left(\mathbf{y}_N^\top \mathbf{y}_N - \mathbf{y}_N^\top \mathbf{X}_N\theta' - (\mathbf{X}_N\theta')^\top \mathbf{y}_N + (\mathbf{X}_N\theta')^\top \mathbf{X}_N\theta'\right) \\
&= \operatorname*{arg\,min}_{\theta' \in \mathbb{R}^{d\times 1}} \left(\mathbf{y}_N^\top \mathbf{y}_N - 2\theta'^\top \mathbf{X}_N^\top \mathbf{y}_N + \theta'^\top \mathbf{X}_N^\top \mathbf{X}_N\theta'\right),
\end{aligned}$$

où la dernière égalité utilise le fait que les quantités $\mathbf{y}_N^\top \mathbf{X}_N\theta'$ et $\theta'^\top \mathbf{X}_N^\top \mathbf{y}_N$ sont égales (étant de dimension $1 \times 1$), ainsi que la propriété $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$ pour deux matrices $\mathbf{A}$ et $\mathbf{B}$. En prenant la dérivée partielle égale à 0 de la dernière équation par rapport à $\theta'$, nous obtenons

$$\begin{aligned}
-2\mathbf{X}_N^\top \mathbf{y}_N + 2\mathbf{X}_N^\top \mathbf{X}_N\theta' &= 0 \\
\mathbf{X}_N^\top \mathbf{X}_N\theta' &= \mathbf{X}_N^\top \mathbf{y}_N,
\end{aligned}$$

donc

$$\theta_N = (\mathbf{X}_N^\top \mathbf{X}_N)^{-1} \mathbf{X}_N^\top \mathbf{y}_N. \tag{1.16}$$

Cependant, pour que ce système puisse être résolu, il importe que la matrice $\mathbf{X}_N^\top \mathbf{X}_N$ soit inversible, donc définie positive. En pratique, cette condition est garantie par l'introduction d'un terme de régularisation. L'estimateur régularisé du paramètre $\theta$ est alors donné par

$$\theta_{\lambda,N} = \left(\mathbf{X}_N^\top \mathbf{X}_N + \lambda \boldsymbol{I}_d\right)^{-1} \mathbf{X}_N^\top \mathbf{y}_N, \tag{1.17}$$

pour un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$, et la prédiction $f_{\lambda,N}(x) = x^\top \theta_{\lambda,N}$.

### 1.2.2 Régression non-paramétrique

Lorsque la fonction $f$ à modéliser n'est pas linéaire, l'idée consiste à utiliser une fonction $\varphi : \mathcal{X} \mapsto \mathcal{K}$ pour projeter les points de l'espace d'entrée $\mathcal{X}$ dans un nouvel espace $\mathcal{K}$, à l'intérieur duquel la relation entre les points projetés et les observations est linéaire. Soit $\varphi(x) = (\varphi_1(x), \dots)^\top$ la projection d'un point $x \in \mathcal{X}$ dans le nouvel espace $\mathcal{K}$, potentiellement de dimension infinie, et soit $\theta = (\theta_1, \dots)^\top$ le paramètre de la fonction linéaire associée dans $\mathcal{K}$, tel que $f(x) = \sum_{i=0}^{\infty} \theta_i \varphi_i(x)$. Par analogie au cas à dimension finie, on dénote $\theta^\top \varphi(x) = \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$. Soit la matrice $N \times \infty$ des points projetés $\boldsymbol{\Phi}_N = (\varphi(x_1), \dots, \varphi(x_N))$. L'estimateur régularisé du paramètre $\theta$ est alors donné par

$$\theta_{\lambda,N} = (\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \lambda \boldsymbol{I}_\infty)^{-1} \boldsymbol{\Phi}_N^\top \mathbf{y}_N, \tag{1.18}$$

pour un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$, et la prédiction $f_{\lambda,N}(x) = \varphi(x)^\top \theta_{\lambda,N}$. Par quelques manipulations algébriques simples, nous avons

$$(\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \lambda \boldsymbol{I}_\infty)\boldsymbol{\Phi}_N^\top = (\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \lambda \boldsymbol{\Phi}_N^\top) = \boldsymbol{\Phi}_N^\top (\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \lambda \boldsymbol{I}_N). \tag{1.19}$$

En multipliant chaque terme de l'égalité par $(\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \lambda \boldsymbol{I}_\infty)^{-1}$ à gauche et $(\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \lambda \boldsymbol{I}_N)^{-1} \mathbf{y}_N$ à droite, nous obtenons

$$\theta_{\lambda,N} = (\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \lambda \boldsymbol{I}_\infty)^{-1} \boldsymbol{\Phi}_N^\top \mathbf{y}_N = \boldsymbol{\Phi}_N^\top (\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \lambda \boldsymbol{I}_N)^{-1} \mathbf{y}_N. \tag{1.20}$$

On remarque que, sous cette forme, $\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top$ est de dimension $N \times N$, contrairement à $\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N$ qui est de dimension $\infty \times \infty$. Cependant, les matrices $\boldsymbol{\Phi}_N$ sont toujours de dimension $N \times \infty$. En pratique, le calcul de vecteurs et de matrices de dimension infinie est problématique et contourné par l'astuce du noyau (*kernel trick*).

#### Méthode à noyau

Soit une fonction noyau (continue, symétrique positive définie) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ sur un ensemble compact [6] $\mathcal{X} \subset \mathbb{R}^d$ doté d'une mesure de Borel positive finie [7]. Alors il existe une séquence dénombrable $(\alpha_i, \psi_i)_{i \in \mathbb{N}^+}$, où $\alpha_i \geqslant 0$, $\lim_{i \to \infty} \alpha_i = 0$ et

$$k(x, x') = \sum_{i=1}^{\infty} \alpha_i \psi_i(x) \psi_i(x'), \quad \forall x, x' \in \mathcal{X}.$$

Soit $\varphi_i(x) = \sqrt{\alpha_i} \psi_i(x)$ . Nous avons donc $k(x, x') = \varphi(x)^\top \varphi(x')$. L'astuce du noyau consiste à remplacer les instances de $\boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top$ et $\varphi(x) \boldsymbol{\Phi}_N^\top$ par le noyau lors du calcul de la prédiction, évitant ainsi d'expliciter des vecteurs et matrices de dimension infinie. Soit la matrice de noyau (*kernel matrix*), ou matrice de Gram,

$$\mathbf{K}_N = \big(k(x_s, x_{s'})\big)_{s,s' \leqslant N}$$

6. Un ensemble compact est borné et fermé. Un ensemble de nombres réels est borné s'il est contenu dans un intervalle fini. Un ensemble est fermé s'il contient tous ses points limitrophes.

7. Une mesure borélienne est une mesure positive définie sur tous les ensembles ouverts du compact.

et le vecteur-colonne de noyau (*kernel vector*)

$$\mathbf{k}_N(x) = \big(k(x, x_s)\big)_{s \leqslant N}$$

associés aux points $\mathbf{X}_N$. Pour un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$, la prédiction suite aux observations $\mathbf{y}_N$ est alors donnée par

$$f_{\lambda,N}(x) = \mathbf{k}_N(x)^\top (\mathbf{K}_N + \lambda \boldsymbol{I}_N)^{-1} \mathbf{y}_N. \tag{1.21}$$

**Complexité de calcul** En pratique, le calcul de la moyenne prédictive requiert l'inversion de la matrice $(\mathbf{K}_N + \lambda \boldsymbol{I}_N) = \mathbf{L}\mathbf{L}^\top$, qui est effectuée par le biais d'une factorisation de Cholesky ainsi que la résolution de deux systèmes linéaires triangulaires.

**Lemme 1** (Factorisation de Cholesky)**.** *Soit une matrice* $\mathbf{A}$ *symétrique définie positive. Il existe alors une matrice triangulaire inférieure* $\mathbf{L}$ *telle que* $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$.

Plus précisément, il s'agit de résoudre le système triangulaire suivant :

$$\begin{aligned} \boldsymbol{\alpha} &= (\mathbf{K}_N + \lambda \boldsymbol{I}_N)^{-1} \mathbf{y}_N \\ (\mathbf{K}_N + \lambda \boldsymbol{I}_N)\boldsymbol{\alpha} &= \mathbf{y}_N \\ \mathbf{L} \underbrace{\mathbf{L}^\top \boldsymbol{\alpha}}_{\boldsymbol{\beta}} &= \mathbf{y}_N, \end{aligned} \tag{1.22}$$

lequel permet d'obtenir $\boldsymbol{\beta}$. Le deuxième système triangulaire,

$$\mathbf{L}^\top \boldsymbol{\alpha} = \boldsymbol{\beta}, \tag{1.23}$$

peut ensuite être résolu pour obtenir $\boldsymbol{\alpha}$. Nous avons finalement

$$f_{\lambda,N}(x) = \mathbf{k}_N(x)^\top \boldsymbol{\alpha}. \tag{1.24}$$

### 1.2.3 Noyaux et RKHS

L'espace de projection $\mathcal{K}$ est connu comme le *reproducing kernel Hilbert space* (RKHS) associé au noyau $k$. On dit que la fonction $f$ appartient à $\mathcal{K}$, si et seulement si $\|f\|_{\mathcal{K}}^2 = \sum_{i=1}^{\infty} \theta_i^2 < \infty$.

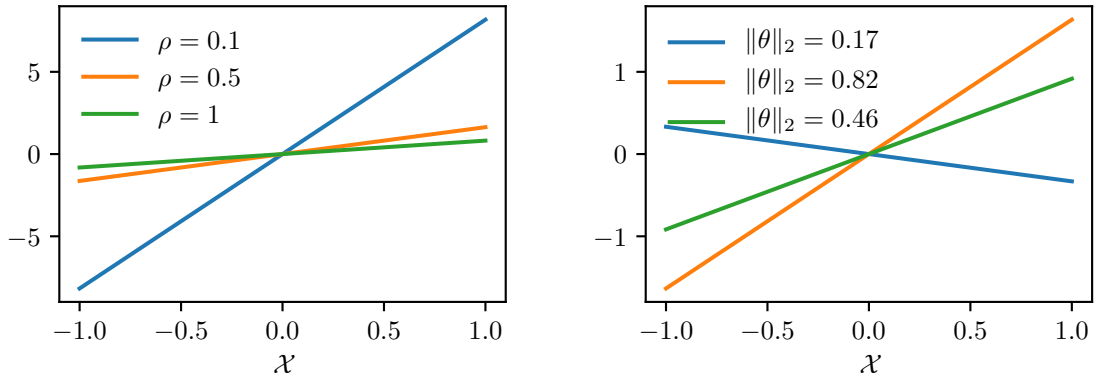
Parmi les noyaux typiques très utilisés dans la littérature, on retrouve le noyau linéaire

$$k(x, x') = \frac{x^\top x'}{\rho^2}, \tag{1.25}$$

permettant de retrouver le modèle paramétrique (équation 1.17), ainsi que le noyau gaussien [8]

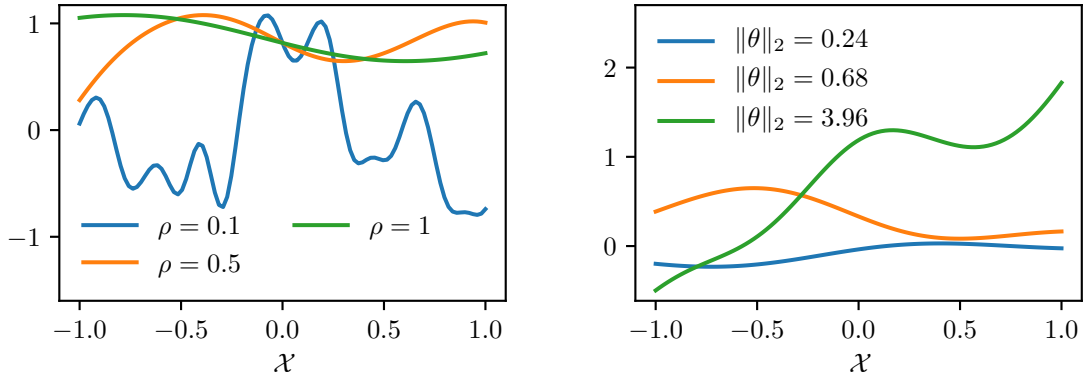
$$k(x, x') = e^{-\frac{(x-x')^2}{2\rho^2}}, \tag{1.26}$$

8. Aussi connu sous le nom de *radial basis function* (RBF) ou de noyau *squared exponential*.




(a) $\|\theta\|_2$ fixe (b) Bande-passante fixe

FIGURE 1.1 – Exemples de fonctions appartenant à des RKHS de noyaux linéaires a) pour différentes bandes-passantes avec $\|\theta\|_2$ fixe et b) pour différents $\|\theta\|_2$ avec bande-passante fixe.




(a) $\|\theta\|_2$ fixe (b) Bande-passante fixe

FIGURE 1.2 – Exemples de fonctions appartenant à des RKHS de noyaux gaussiens a) pour différentes bandes-passantes avec $\|\theta\|_2$ fixe et b) pour différents $\|\theta\|_2$ avec bande-passante fixe.

dans lesquels un paramètre de bande-passante (*bandwidth* ou *length scale*) $\rho$ décrit la régularité des fonctions supportées par le RKHS résultant. Par exemple, une bande-passante plus courte indique un plus faible transfert d'information entre deux points, donc une variation potentielle plus importante d'une fonction entre ces deux points. Cela est illustré par les figures 1.1a et 1.2a montrant les fonctions de norme fixe obtenues dans les RKHS correspondant aux noyaux linéaires et gaussiens de différentes bandes-passantes. La régularité d'une fonction appartenant à un RKHS donné dépend également de la norme de la fonction pour un noyau donné : plus la norme est élevée, plus la fonction admet de variabilités. Cela est illustré par les figures 1.1b et 1.2b montrant des fonctions de normes différentes appartenant au RKHS de noyaux linéaires et gaussiens de bande-passante $\rho = 0.5$.

### 1.2.4 Modèle de régression a posteriori

Considérons des observations bruitées telles que $y = f(x) + \xi$, avec $\xi \sim \mathcal{N}(0, \sigma^2)$ et un a priori $\theta \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\theta)$. La vraisemblance des observations $\mathbf{y}_N$ est donc donnée par

$$\mathbb{P}[y_1, \dots, y_N | x_1, \dots, x_N, f] \sim \mathcal{N}\Big(\Big[f(x_1), \dots, f(x_N)\Big], \sigma^2 \boldsymbol{I}_N\Big)$$

et la distribution postérieure prédictive de la fonction évaluée en un point $x$ est donnée par

$$\widehat{f}_N(x) | x_1, \dots, x_N, y_1, \dots, y_N \sim \mathcal{N}(f_N(x), s_N^2(x)),$$

avec

$$\begin{aligned}
f_N(x) &= \varphi(x)^\top (\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \sigma^2 \boldsymbol{\Sigma}_\theta^{-1})^{-1} \boldsymbol{\Phi}_N^\top \mathbf{y}_N, && (1.27)\\
&= \varphi(x)^\top \boldsymbol{\Sigma}_\theta (\boldsymbol{\Phi}_N^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}_N + \sigma^2 \boldsymbol{I}_\infty)^{-1} \boldsymbol{\Phi}_N^\top \mathbf{y}_N, && (1.28)\\
&= \varphi(x)^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}_N^\top (\boldsymbol{\Phi}_N \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}_N^\top + \sigma^2 \boldsymbol{I}_N)^{-1} \mathbf{y}_N \quad \text{et} && (1.29)\\
s_N^2(x) &= \sigma^2 \varphi(x)^\top (\boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N + \sigma^2 \boldsymbol{\Sigma}_\theta^{-1})^{-1} \varphi(x) && (1.30)\\
&= \varphi(x)^\top \boldsymbol{\Sigma}_\theta \varphi(x) - \varphi(x)^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}_N^\top (\boldsymbol{\Phi}_N \boldsymbol{\Sigma}_\theta \boldsymbol{\Phi}_N^\top + \sigma^2 \boldsymbol{I}_N)^{-1} \boldsymbol{\Phi}_N \boldsymbol{\Sigma}_\theta \varphi(x), && (1.31)
\end{aligned}$$

où l'égalité 1.31 est obtenue avec la formule de Sherman-Morrison. Cette formulation, dénotée *Gaussian process* (GP), peut être vue comme une généralisation de la distribution de probabilité normale de l'espace des variables à l'espace des fonctions, dans laquelle un processus stochastique gouverne les propriétés d'une distribution normale à chaque point de l'espace $\mathcal{X}$ (Rasmussen and Williams, 2006).

Ce résultat permet d'interpréter le paramètre de régularisation $\lambda$ comme un a priori sur la variance (Maillard, 2016), $\boldsymbol{\Sigma}_\theta = \frac{\sigma^2}{\lambda} \boldsymbol{I}_\infty$, et de généraliser l'espérance et la variance prédictive par

$$\begin{aligned}
f_{\lambda,N}(x) &= \mathbf{k}_N(x)^\top (\mathbf{K}_N + \lambda \boldsymbol{I}_N)^{-1} \mathbf{y}_N \quad \text{et} && (1.32)\\
s_{\lambda,N}^2(x) &= \frac{\sigma^2}{\lambda} k_{\lambda,N}(x,x) \quad \text{où} \quad k_{\lambda,N}(x,x') = k(x,x') - \mathbf{k}_N(x)^\top (\mathbf{K}_N + \lambda \boldsymbol{I}_N)^{-1} \mathbf{k}_N(x'). && (1.33)
\end{aligned}$$

La distribution jointe sur les observations antérieures et l'évaluation en un point $x$ correspond donc à une distribution normale multivariée

$$\begin{bmatrix} \mathbf{y}_N \\ f(x) \end{bmatrix} \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} \begin{bmatrix} \mathbf{K}_N + \lambda \boldsymbol{I}_N & \mathbf{k}_N(x) \\ \mathbf{k}_N(x)^\top & k(x,x) \end{bmatrix}\right)$$

et la distribution postérieure sur $f$ est donnée par

$$\mathbb{P}[f | x_1, \dots, x_N, y_1, \dots, y_N] \sim \mathcal{N}\Big(\Big(f_{\lambda,N}(x)\Big)_{x \in \mathcal{X}}, \frac{\sigma^2}{\lambda}\Big[k_{\lambda,N}(x,x')\Big]_{x,x' \in \mathcal{X}}\Big).$$

### 1.2.5 Gain d'information et régularisation

Dans la situation où l'acquisition du prochain point $N+1$ se base sur le modèle de régression conditionné sur les $N$ observations précédentes, le *gain d'information* peut constituer une mesure pertinente pour l'analyse de la convergence de l'estimateur par régression. Cette quantité mesure l'information recueillie sur une fonction $f$ en échantillonnant aux points $(x_1, \ldots, x_N)$. Elle est définie comme l'*information mutuelle* entre la fonction sous-jacente et l'ensemble des observations $(y_1, \ldots, y_N)$ en ces points :

$$I(y_1, \ldots, y_N; f) = H(y_1, \ldots, y_N) - H(y_1, \ldots, y_N | f),$$

soit la différence entre l'*entropie marginale* et l'*entropie conditionnelle* des distributions d'observations. Rappelons que l'entropie correspond à la quantité d'information fournie en moyenne par une source de données stochastique. Le gain d'information quantifie donc la réduction de l'incertitude sur $f$ suite à l'observations de $y_1, \ldots, y_N$. Un gain d'information plus important se produit lorsque des points sont échantillonnés dans une région méconnue de la fonction. En contrôlant le gain d'information, il est possible de contrôler la quantité d'échantillons recueillis dans les régions possiblement sous-optimales.

Pour une normale multidimensionnelle, $H(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2} \ln |2\pi e \boldsymbol{\Sigma}|$, de telle sorte que

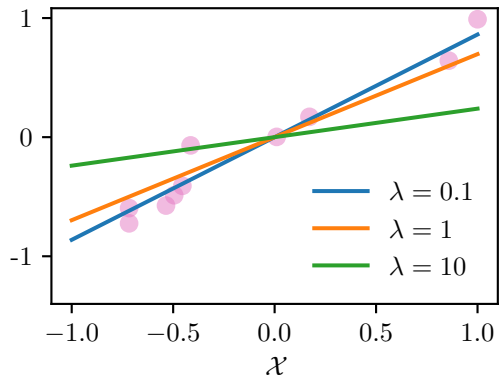
$$I(y_1, \ldots, y_N; f) = \frac{1}{2} \ln |\boldsymbol{I}_N + \sigma^{-2} \mathbf{K}_N|$$

dans un GP (Srinivas et al., 2010). Nous généralisons ce concept à la régression à noyau de régularisation $\lambda$ quelconque.

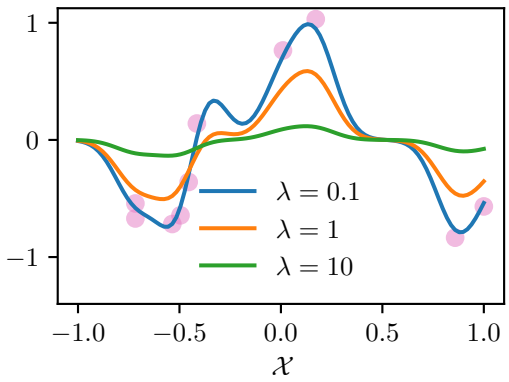
**Définition 5** (Gain d'information avec bruit inconnu)**.** *Nous définissons le gain d'information avec $N$ observations pour un paramètre de régularisation $\lambda$ comme*

$$\gamma_N(\lambda) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{s=1}^{N} \ln \Big(1 + \frac{1}{\lambda} k_{\lambda, s-1}(x_s, x_s)\Big).$$

Le gain d'information est inversement proportionnel au paramètre de régularisation $\lambda$. En contrôlant la variabilité tolérée de la fonction estimée, le paramètre de régularisation limite l'impact d'une observation sur le modèle résultant. La figure 1.3 montre l'espérance prédictive obtenue par des modèles de régression à noyau (équation 1.32) de différentes régularisations, conditionnés sur les observations données. On voit bien qu'une régularisation plus élevée correspond à une agglomération de l'information dans l'espace. Plus la régularisation est importante, plus le transfert de connaissance des observations $(y_1, \ldots, y_N)$ vers le point $x_{N+1}$ est important, donc moins l'acquisition d'une observation $y_{N+1}$ apporte d'information *nouvelle*. Ainsi, il est naturel qu'une régularisation plus importante implique un gain d'information plus faible. La figure 1.4 montre la courbe de croissance du gain d'information en fonction du nombre d'observations (i.i.d.), pour différentes régularisations. Tel qu'attendu, la convergence du gain d'information est inversement proportionnelle à l'amplitude de la régularisation.

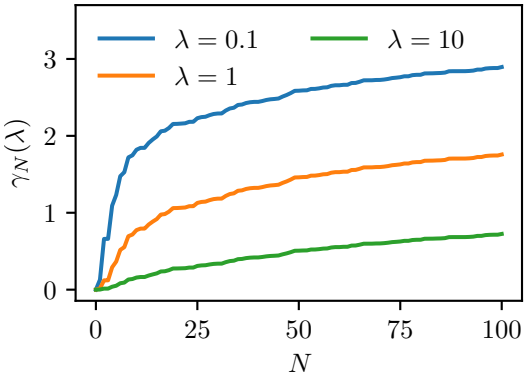



(a) Noyau linéaire ($\rho = 1$)

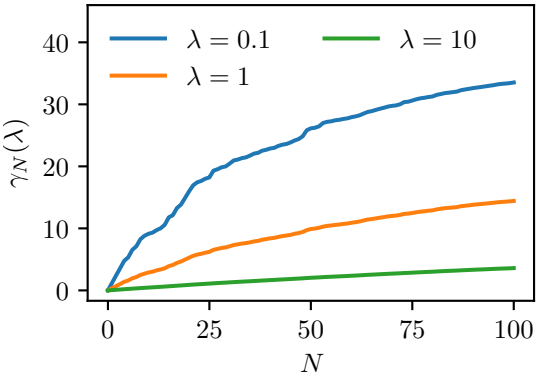



(b) Noyau gaussien ($\rho = 0.1$)

FIGURE 1.3 – Impact de la régularisation sur l'espérance prédictive (équation 1.32) obtenue par la régression à noyau sur 10 observations (rose).




(a) Noyau linéaire ($\rho = 1$)




(b) Noyau gaussien ($\rho = 0.1$)

FIGURE 1.4 – Croissance du gain d'information (définition 5) en fonction du nombre d'observations, pour différents paramètres de régularisation.

Finalement, il faut noter que le gain d'information est lié à la dimensionnalité effective du problème (Valko et al., 2013). En effet, les dimensions *non effectives* sont, par définition, non informatives. Si un nombre d'observations plus important est nécessaire pour bien couvrir l'espace, cela signifie que plus d'observations sont porteuses d'information sur la fonction. Cela se traduit donc directement par une convergence plus lente du gain d'information.

## 1.3 La régression pour l'apprentissage en-ligne

Les approches de régression ont teinté les adaptations des algorithmes UCB (Srinivas et al., 2010; Abbasi-Yadkori et al., 2011; Krause and Ong, 2011; Valko et al., 2013) et TS (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016) dans diverses variantes de bandits, tels les bandits contextuels et les bandits structurés traités aux chapitres suivants. Les estimateurs

par régression permettent de généraliser les algorithmes traditionnels (voir la section 1.1.2), visant à estimer des espérances, à l'estimation de fonctions d'espérances définies sur un espace d'entrée structuré donné, soit les contextes ou les actions.

La convergence des algorithmes de bandits basés sur la régression repose donc sur les garanties de convergence des estimateurs par régression. La concentration des estimateurs par régression utilisés conjointement à des approches de bandits présente certains défis puisque les observations ne sont pas i.i.d. En effet, le modèle de régression, conditionné sur les observations antérieures, sert à la prise de décision pour l'acquisition d'observations subséquentes. Les observations étant ainsi dépendantes du passé, des techniques d'analyse particulières (Abbasi-Yadkori et al., 2011; Maillard, 2016) doivent être considérées pour obtenir les garanties théoriques appropriées.

# Chapitre 2

# Les bandits contextuels

Dans ce chapitre, nous présentons une étude de cas spécifique, où nous visons à concevoir des stratégies efficaces d'allocation de traitement pour des essais cliniques adaptatifs afin d'éventuellement optimiser les stratégies de traitement pharmacologique personnalisées pour le cancer. Un ensemble de données initial a été collecté avec une procédure d'allocation aléatoire. Ce processus d'acquisition de données est coûteux et prend beaucoup de temps. Nous visons donc à concevoir une stratégie d'allocation de traitement adaptative pour améliorer l'efficacité de la collecte de données dans la prochaine phase d'expériences.

Nous formalisons cette application comme un problème de bandits contextuels et nous l'abordons avec un algorithme basé sur la régression à noyau et le sous-échantillonnage : GP BESA. Des expériences simulées mises en place à partir des données initiales et appuyant l'intuition derrière GP BESA servent ensuite de motivation à la réalisation de la seconde phase d'expériences sur des données animales.

## 2.1 Formulation du problème

Un problème de bandits contextuels est décrit par un ensemble de contextes $\mathcal{X}$ et un ensemble d'actions $\mathcal{A}$. Dans une application de médecine de précision, le contexte peut inclure des caractéristiques du sujet et de la maladie tandis que les actions peuvent représenter différents traitements ou stratégies de prévention disponibles. Nous considérons ici le cas où $\mathcal{X}$ est compact et $\mathcal{A}$ est fini. Plus spécifiquement, nous considérons la situation où les actions sont catégoriques (disjointes), c'est-à-dire où il n'y a aucune similarité entre elles. Cela représente une situation classique dans laquelle il est difficile, voire impossible, d'établir une métrique de similarité dans l'espace des traitements. Chaque action $a \in \mathcal{A}$ est associée à une fonction inconnue $f_a : \mathcal{X} \mapsto \mathbb{R}$. À chaque épisode $t \in \mathbb{N}_{>0}$, un agent observe le contexte $x_t$, choisit d'effectuer une action $a_t$ et obtient donc une observation $y_t$ perturbée par un bruit $\xi_t$, telle que $y_t := f_{a_t}(x_t) + \xi_t$.

Un algorithme de bandits contextuels est une méthode (possiblement randomisée) pour sélectionner la prochaine action à effectuer étant donné l'historique des contextes observés précédemment, des actions antérieures et des observations obtenues, $\mathcal{H}_t := \{x_s, a_s, y_s\}_{s=1}^t$. Soit l'action optimale $\star_t := \arg\max_{a\in\mathcal{A}} f_a(x_t)$. Le but de l'algorithme est de minimiser le pseudo-regret cumulatif :

$$\mathfrak{R}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \big[f_{\star_t}(x_t) - f_{a_t}(x_t)\big]. \tag{2.1}$$

Cette quantité mesure la performance de l'algorithme comparée à celle d'un oracle qui connaîtrait les fonctions de récompenses $f_a$ pour tout $a \in \mathcal{A}$.

**Remarque 2.** *Le problème de bandits traditionnel peut être formulé comme un problème de bandits contextuels dans lequel 1)* $f_a(x) = \mu_a$ *pour tout* $x \in \mathcal{X}$, $a \in \mathcal{A}$ *ou 2)* $|\mathcal{X}| = 1$.

**Hypothèses** Pour la suite, nous considérons un bruit gaussien, $\xi_t \sim \mathcal{N}(0, \sigma^2)$.

## 2.2 Littérature

Le problème des bandits contextuels a été abordé par le biais de différentes hypothèses de similarité entre les actions et les contextes. Plusieurs travaux (Auer et al., 2002a; Li et al., 2010; Chu et al., 2011; Agrawal and Goyal, 2014; Abeille and Lazaric, 2016) font l'hypothèse que les fonctions $f_a$ sont linéaires sur $\mathcal{X}$. D'autres abordent la situation dans laquelle l'ensemble des actions $\mathcal{A}$ est compact avec une mesure de similarité donnée. Ils considèrent donc une fonction d'observations $f : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ telle que $f_a(x) = f(x, a)$. Parmi ceux-ci, Slivkins (2014) fait l'hypothèse que $f$ est Lipschitz, tandis que Krause and Ong (2011) et Valko et al. (2013) font l'hypothèse que $f$ est échantillonnée d'un *Gaussian process* (GP) ou contenue dans un *reproducing kernel Hilbert space* (RKHS). Plus spécifiquement, Krause and Ong (2011) introduisent l'algorithme CGP-UCB qui sélectionne l'action

$$a_t = \arg\max_{a\in\mathcal{A}} \Big[f_{t-1}(a, x_t) + B_t(\delta) s_{t-1}(a, x_t)\Big], \tag{2.2}$$

où $f_{t-1}(\cdot)$ et $s_{t-1}(\cdot)$ représentent respectivement la moyenne et l'écart type a posteriori d'un GP sur l'espace joint $\mathcal{X} \times \mathcal{A}$ (voir les équations 1.27 et 1.30), et le second terme correspond à un intervalle de confiance sur l'estimateur $f_{t-1}(a, x_t)$ tel que

$$\mathbb{P}\Big[|f_{t-1}(a, x_t) - f_a(x_t)| > B_t(\delta) s_{t-1}(a, x_t)\Big] \leqslant \delta$$

pour tout $\delta \in [0, 1]$. L'algorithme KernelUCB (Valko et al., 2013) généralise CGP-UCB à un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$ quelconque et sélectionne l'action

$$a_t = \arg\max_{a\in\mathcal{A}} \left[f_{\lambda,t-1}(a, x_t) + B_{\lambda,t}(\delta) \sqrt{\frac{k_{\lambda,t-1}((a, x_t), (a, x_t))}{\lambda}}\right], \tag{2.3}$$

où $f_{\lambda,t-1}(\cdot)$ et $k_{\lambda,t-1}(\cdot)$ sont respectivement données par les équations 1.32 et 1.33 appliquées sur l'espace joint $\mathcal{X} \times \mathcal{A}$, et le second terme correspond à un intervalle de confiance sur l'estimateur $f_{\lambda,t-1}(a, x_t)$ tel que

$$\mathbb{P}\left[|f_{\lambda,t-1}(a,x_t) - f_a(x_t)| > B_{\lambda,t}(\delta)\sqrt{\frac{k_{\lambda,t-1}((a,x_t),(a,x_t))}{\lambda}}\right] \leqslant \delta$$

pour tout $\delta \in [0,1]$. Pour $\lambda = \sigma^2$, KernelUCB et CGP-UCB sont équivalents.

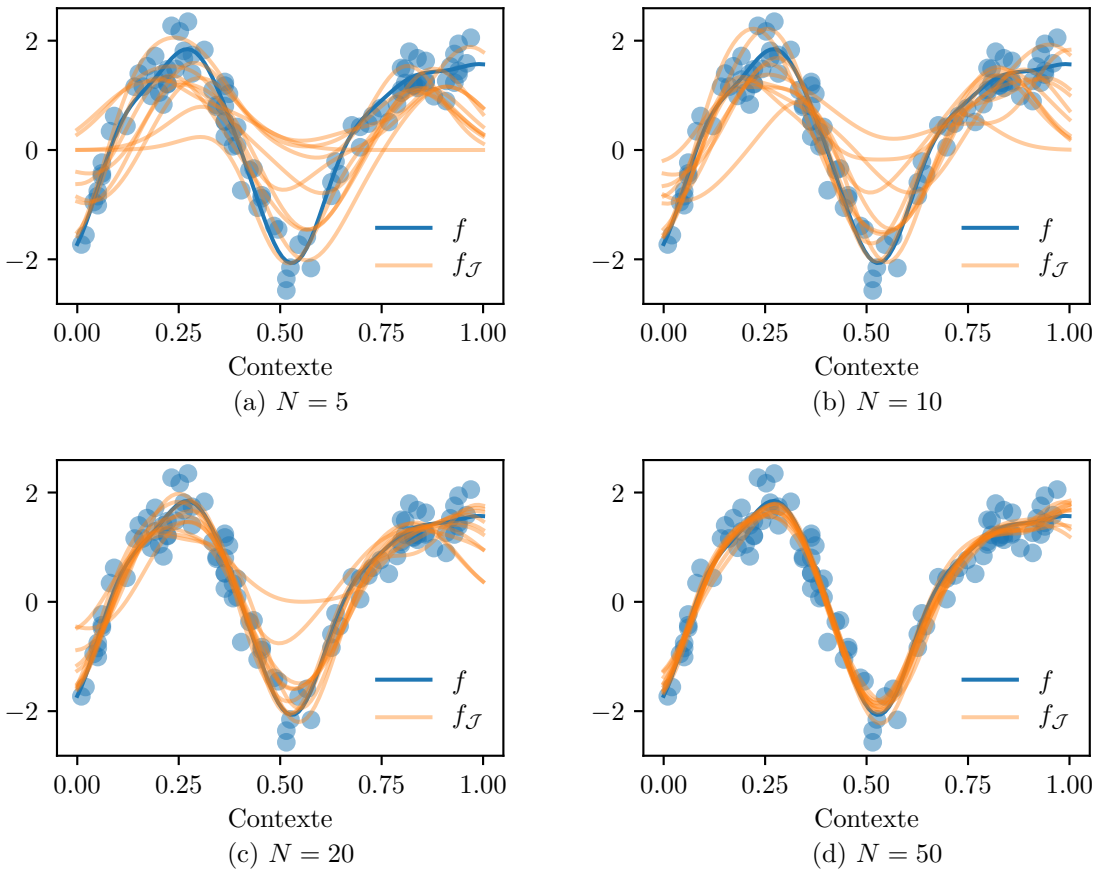
Sous l'hypothèse que chaque fonction $f_a$ provient d'un GP indépendant, le problème abordé dans ce chapitre devient un cas spécifique du problème abordé par Krause and Ong (2011) et Valko et al. (2013) dans lequel il n'y a aucune similarité entre les actions. Cela peut être considéré comme une *distance infinie* entre les actions, laquelle peut être représentée par une bande-passante $\rho \to 0$ (voir la section 1.2.3). Cette variante est connue dans la littérature comme le problème des bandits à covariables (*bandits with covariates*) (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013).

Il faut également noter que le problème des bandits contextuels peut être vu comme une instance spécifique des *Markov decision process*s (MDPs) en *reinforcement learning* (RL) (Ghavamzadeh et al., 2015). Un MDP est décrit par un ensemble d'états (correspondant aux contextes ici), un ensemble d'actions, ainsi qu'une matrice des probabilités de transition entre les états étant donné les différentes actions. Un problème de bandits contextuels peut être vu comme un MDP dans lequel la transition entre les états (c'est-à-dire l'arrivée des contextes) est indépendante des actions entreprises.

## 2.3 GP BESA

Dans le problème des bandits contextuels avec des actions catégoriques, il est possible que les fonctions $f_a$ soient de régularités différentes. En effet, il est réaliste de croire que les effets de différents traitements peuvent varier plus ou moins brusquement étant donné le volume d'une tumeur. Dans ce cas, plutôt que de s'appuyer sur un modèle de régression commun, donc sur un noyau de bande-passante commune entre les actions, il est naturel de modéliser chaque fonction $f_a$ de manière indépendante en utilisant son propre noyau.

Nous introduisons GP BESA, une généralisation de l'algorithme *best empirical sampled average* (BESA) (voir la section 1.1) au problème de bandits contextuels avec des actions catégoriques. L'idée consiste à modéliser la fonction des observations $f_a$ pour chaque action $a \in \mathcal{A}$ en utilisant un estimateur de régression GP conditionné sur un sous-échantillon des observations disponibles, de taille équivalente pour chaque action. Rappelons que l'algorithme BESA traditionnel, ne repose que sur l'estimateur empirique de la moyenne pour sélectionner la prochaine action. En effet, la variance dans les observations est représentée par la variance dans l'estimateur de la moyenne empirique étant donné différents sous-échantillonnages. De la




(a) $N = 5$ (b) $N = 10$

(c) $N = 20$ (d) $N = 50$

FIGURE 2.1 – Exemples de moyennes prédites par un GP conditionné sur 10 sous-échantillons différents $\mathcal{J}$ de taille $N$ parmi 100.

même manière, GP BESA ne repose que sur la variance introduite par le sous-échantillonnage dans l'estimateur de la moyenne postérieure, sans égard à la variance postérieure.

Soit un ensemble d'épisodes $\mathcal{J}$ et soit les observations associées $\mathbf{y}_{\mathcal{J}} = (y_j)_{j\in\mathcal{J}}^{\top} \in \mathbb{R}^{|\mathcal{J}|}$. Alors pour un noyau $k(\cdot,\cdot)$, avec

$$\mathbf{K}_{\mathcal{J}} = \big(k(x_j, x_{j'})\big)_{j,j'\in\mathcal{J}} \quad \text{et} \quad \mathbf{k}_{\mathcal{J}}(x) = \big(k(x, x_j)\big)_{j\in\mathcal{J}},$$

on dénote l'espérance et la variance postérieure prédictive en $x$ étant données les observations obtenues aux épisodes $\mathcal{J}$ comme

$$f_{\mathcal{J}}(x) = \mathbf{k}_{\mathcal{J}}(x)^{\top}(\mathbf{K}_{\mathcal{J}} + \sigma^2 \boldsymbol{I}_{|\mathcal{J}|})^{-1}\mathbf{y}_{\mathcal{J}} \quad \text{et} \tag{2.4}$$

$$s^2_{\mathcal{J}}(x) = k(x,x) - \mathbf{k}_{\mathcal{J}}(x)^{\top}(\mathbf{K}_{\mathcal{J}} + \sigma^2 \boldsymbol{I}_{|\mathcal{J}|})^{-1}\mathbf{k}_{\mathcal{J}}(x). \tag{2.5}$$

Soit $\mathcal{J}_{a,t} := \{j\}_{1\leqslant j\leqslant t, a_j = a}$ l'ensemble des épisodes antérieurs lors desquels l'action $a$ a été sélectionnée jusqu'au temps $t$ (inclusivement) et soit le nombre d'observations correspondantes $N_{a,t} = |\mathcal{J}_{a,t}|$. La figure 2.1 montre des exemples de moyennes prédictives obtenues par régres-

**Algorithme 5** Sélection GP BESA pour deux actions.

Paramètres : épisode en cours $t$, contexte $x_t$, deux actions $a$ et $b$

1: $N = \min(N_{a,t-1}, N_{b,t-1})$
2: $\mathcal{I}_a \leftarrow \texttt{subsample}(N, N_{a,t-1})$ et $\mathcal{I}_b \leftarrow \texttt{subsample}(N, N_{b,t-1})$
3: calculer $\tilde{f}_{a,t} = f_{\mathcal{J}_{a,t-1}(\mathcal{I}_a)}(x_t)$ et $\tilde{f}_{b,t} = f_{\mathcal{J}_{b,t-1}(\mathcal{I}_b)}(x_t)$
4: $a_t = \arg\max_{i \in \{a,b\}} \tilde{f}_{i,t}(x_t)$
5: **return** $a_t$

sion GP sur des sous-échantillons de différentes tailles parmi un bassin de 100 échantillons générés par la fonction de référence (bleu) soumise à un bruit d'écart type $\sigma = 0.3$.

**Remarque 3.** *L'espérance postérieure $f_{\mathcal{J}_{a,t}}(x)$ donnée par l'équation 2.4 correspond à l'espérance postérieure $f_t(x)$ donnée par l'équation 1.27, ce qui correspond également à l'espérance postérieure $f_{\sigma^2,t}(x)$ donnée par l'équation 1.32.*

Nous adoptons la convention $\mathcal{J}_{a,t}(\mathcal{I}) := \{j_{i_1}, \ldots, j_{i_m}\}$ pour des indices $\mathcal{I} = \{i_1, \ldots, i_m\}$ et $i \leqslant N_{a,t}$ pour tout $i \in \mathcal{I}$. L'algorithme 5 décrit la procédure pour sélectionner l'action à effectuer au temps $t$ avec GP BESA pour le cas de deux actions ; tout comme BESA, il est facilement extensible à de multiples actions en organisant des tournois par paire d'actions (voir l'algorithme 4 à la section 1.1.2). En calculant le modèle de régression GP sur le même nombre d'observations pour chaque action, GP BESA vise à obtenir des modèles de régression d'intervalles de confiance comparables, suivant la philosophie de BESA. Similairement à BESA, la composante explorative de GP BESA provient du sous-échantillon sélectionné aléatoirement pour évaluer l'estimateur. Contrairement aux approches existantes (Krause and Ong, 2011; Valko et al., 2013), GP BESA est aléatoire. Cette caractéristique sera déterminante pour l'application présentée à la section suivante.

**Complexité** D'un point de vue computationnel, GP BESA doit conserver en mémoire l'ensemble des observations obtenues avec chaque action et doit sous-échantillonner au plus $|\mathcal{A}|-1$ fois lors des tournois. Les expériences de Baransi et al. (2014) avec BESA concluent que cela n'est pas problématique. GP BESA doit en outre calculer la moyenne prédictive sur la fonction $f_a$ pour chaque action $a \in \mathcal{A}$, à chaque étape d'un tournoi. Cela correspond au plus à $2^{\lceil \log_2 |\mathcal{A}| \rceil + 1} - 2$ répétitions, chacune incluant l'évaluation du noyau sur les points sous-échantillonnés et la résolution de deux systèmes d'équations linéaires triangulaires (voir la section 1.2.2). L'évaluation du noyau requiert de calculer la distance entre chacune des paires de points sous-échantillonnés. Ce processus peut être accéléré par la mise en cache de ces valeurs pour éviter leur recalcul. Les systèmes d'équations linéaires triangulaires doivent cependant être résolus à chaque fois.

## 2.4 Allocation de traitements adaptative

La médecine de précision a le potentiel d'améliorer significativement la réponse aux traitements en offrant des stratégies de traitements adaptées aux patients et à la maladie. Nous considérons ici une étude sur des souris avec des tumeurs cancéreuses induites (Balmain et al., 1984) que nous envisageons de traiter par des combinaisons de 5-FU, un agent chimiothérapeutique, et d'imiquimod, un composé synthétique modifiant la réponse immunitaire. Nous considérons ainsi les options suivantes, à dosage fixe : aucun traitement, 5-FU, imiquimod et combinaison simultanée d'imiquimod et de 5-FU. Plus précisément, nous aimerions apprendre des politiques de traitements qui s'adaptent au stade de la maladie. À l'heure actuelle, la seule information que nous possédons sur le stade du cancer réside dans les mesures tumorales. Par conséquent, nous caractérisons la maladie en utilisant le volume tumoral $v$ calculé par

$$v = \frac{\pi}{6}(\ell w)^{3/2},$$

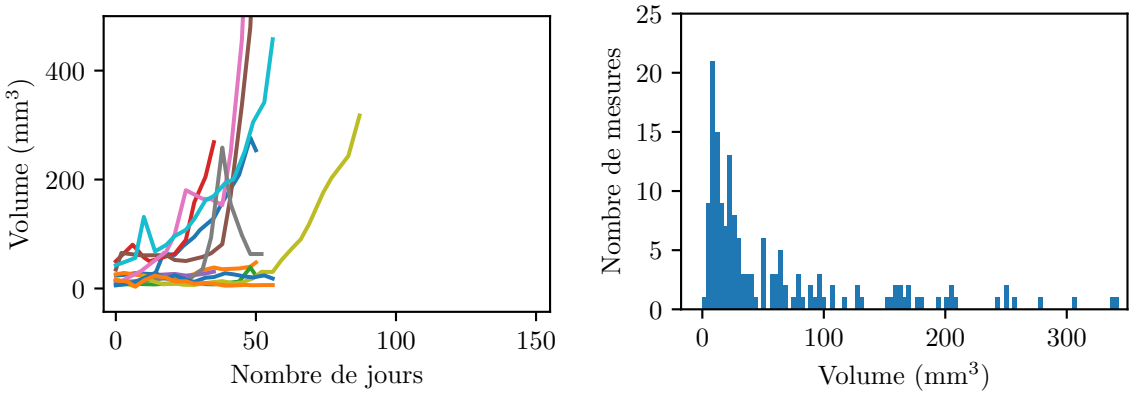
où $\ell$ et $w$ dénotent respectivement la longueur et la largeur de la tumeur éllipsoïdale (Tomayko and Reynolds, 1989).

Dans un *randomized clinical trial* (RCT) classique, les traitements disponibles sont assignés aléatoirement à des sujets, indépendamment de leurs caractéristiques et des caractéristiques de la maladie. Ce processus de cueillette de données peut s'avérer coûteux en temps et en ressources. L'étude considérée ici a débuté par un RCT. Lors de cette phase initiale, des traitements ont été assignés aléatoirement deux fois par semaine à six souris ayant développé jusqu'à trois tumeurs chacune. L'évolution d'une tumeur est représentée par une série de triplets $(v_i, a_i, v_i')$, où $v_i$ dénote la $i$-ème mesure du volume de la tumeur, $a_i$ est le traitement assigné le jour de la mesure $v_i$ et $v_i'$ est le volume tumoral mesuré après l'effet du traitement. L'objectif est d'apprendre, pour chaque traitement $a$, la fonction de transition entre $v_i$ et $v_i'$ en utilisant tous les triplets où $a_i = a$. Après le retrait des données non utilisables pour des raisons relevant de l'aléatoire biologique, par exemple, des tumeurs non développées ou à croissance démesurée pour cause de fusion, un total de 163 triplets provenant de cinq souris (12 tumeurs) sont disponibles initialement. Le tableau 2.1 montre comment ces données sont réparties entre les différentes options de traitements. Les figures 2.2a et 2.2b montrent respectivement la croissance des tumeurs au fil des jours ainsi que la répartition des volumes tumoraux observés durant cette phase. On observe une croissance souvent exponentielle des tumeurs, pouvant être attribuable à l'assignation de traitements peu efficaces. Cette situation entraîne une dégradation rapide des sujets, limitant le nombre de données pouvant être recueillies par sujet, et restreint également l'espace des états (volumes de tumeur) visités. Cela pourrait se révéler problématique dans l'utilisation ultérieure des données pour établir une politique personnalisée.

Cela motive la mise en place d'un *adaptive clinical trial* (ACT), dans lequel les données accumulées sur les réponses aux traitements sont utilisées pour favoriser l'allocation de *meilleurs*

Tableau 2.1 – Nombre d'échantillons par traitement dans l'ensemble de données initial.

| | Aucun | 5-FU | Imiquimod | 5-FU + Imiquimod | Total |
|---|---|---|---|---|---|
| $N$ | 42 | 66 | 24 | 31 | 163 |




(a) Évolution des tumeurs

(b) Distribution des volumes de tumeur

FIGURE 2.2 – Données initiales acquises durant le RCT.

traitements, réduisant ainsi l'exposition aux traitements moins efficaces. Cette approche donne lieu au fameux compromis entre l'exploration (permettant de découvrir le traitement optimal) et l'exploitation (consistant à traiter les patients le plus efficacement possible), ce qui en fait un cadre applicatif de choix pour les algorithmes de bandits. Plus spécifiquement, l'ACT peut être formulé comme un problème de prise de décision séquentielle, dans lequel il importe de choisir une séquence d'actions (traitements) permettant d'obtenir un résultat donné, typiquement la guérison du patient. Ce problème est abordé avec des approches de RL, soit les MDPs (Villar et al., 2015) [1]. Ces dernières permettent de gérer non seulement la réponse immédiate à un traitement, mais également l'impact de la séquence de traitements sur le résultat final, c'est-à-dire la guérison du patient. La résolution d'un MDP basée sur les équations de Bellman (1952) ou sur les indexes de Gittins (1974) requiert cependant assez de données pour bien couvrir l'espace des états. Cela n'est pas le cas dans la présente situation, où les données acquises en RCT sont très concentrées dans les petits volumes de tumeurs. De plus, pour des raisons de logistique, l'ACT dans l'étude considérée ici se déroule sur des groupes d'animaux et l'historique des observations utilisé par l'algorithme d'allocation n'est mis à jour qu'après l'achèvement d'un groupe donné. Ainsi, nous faisons face à un problème avec rétroactions retardées (*delayed feedback*). L'algorithme 6 montre la procédure résultante. Les algorithmes d'ACT basés sur les MDPs sont actuellement déterministes et, par conséquent, peu efficaces dans un environnement à rétroactions retardées (Williamson et al., 2017). Les algorithmes randomisés sont connus pour être plus robustes à l'obtention des rétroactions avec retard (Chapelle and Li, 2011).

1. Aussi connu sous le terme *Bayesian Bernoulli multi-armed bandit problem.*

**Algorithme 6** Procédure d'ACT par groupes d'animaux (avec rétroactions retardées).

$t \leftarrow 0$
**for all** groupe $g$ **do**
  initialiser l'historique du groupe $\mathcal{D}_g$
  **for all** souris dans le groupe $g$ **do**
    **repeat**
      observer le volume de tumeur $x_t$
      sélectionner le traitement $a_t$ (ici avec GP BESA) et l'administrer
      observer l'effet du traitement $x'_t$, avec $y_t = -x'_t$
      ajouter le tuple $(x_t, a_t, y_t)$ à l'historique $\mathcal{D}_g$
      $t \leftarrow t + 1$
    **until** tumeur trop volumineuse
  **end for**
  **for all** $(x_\tau, a_\tau, y_\tau) \in \mathcal{D}_g$ **do**
    mettre à jour l'historique des épisodes
  **end for**
**end for**

Pour ces raisons, nous choisissons d'aborder ce problème dans le cadre des bandits contextuels, où nous considérons l'espace des contextes $\mathcal{X}$ correspondant à l'espace des volumes tumoraux, ainsi que l'ensemble d'actions $\mathcal{A}$ correspondant aux options de traitements. Chaque épisode $t$ est décrit par l'observation d'un volume de tumeur $x_t$, l'assignation d'un traitement $a_t$ et l'observation du volume subséquent $x'_t$. Ce dernier se traduit en récompense $y_t = -x'_t$ puisque les récompenses sont maximisées. Lorsqu'un groupe d'animaux arrive à terme, tous les triples obtenus pour ce groupe sont utilisés pour mettre à jour l'algorithme de recommandation de traitements. On peut considérer cette formulation comme une approximation myope d'un problème de RL dans laquelle nous visons à optimiser uniquement la valeur de l'état suivant, un état étant décrit par le volume de la tumeur. Un algorithme pour ce problème vise à réduire le volume des tumeurs, ce qui devrait intuitivement permettre de ralentir la croissance des tumeurs. Cela devrait permettre de recueillir des échantillons à plus long terme en étirant la durée de vie des sujets animaux, tout en visitant des états plus divers en retardant la croissance exponentielle des tumeurs.

## 2.5 Expériences en simulation

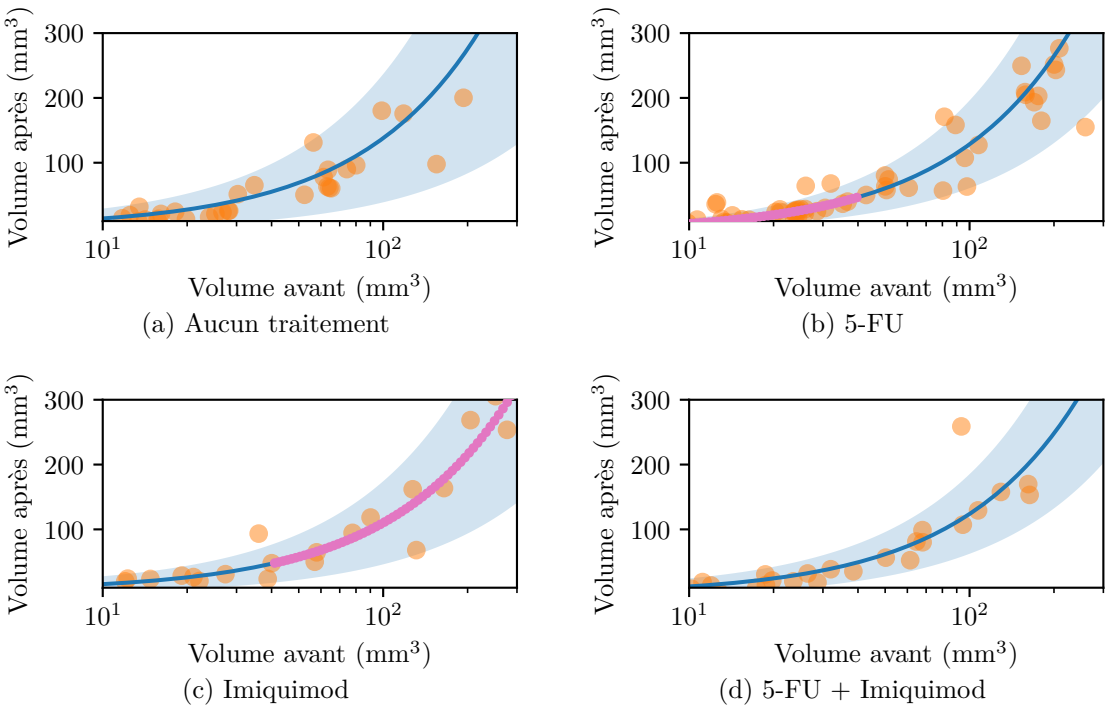
Nous effectuons des expériences en simulation afin d'évaluer la performance de la GP BESA avant de déployer cette stratégie sur de vrais animaux. Les modèles de simulation sont construits à partir des données disponibles recueillies lors de la phase de RCT initiale décrite à la section 2.4. Nous modélisons l'arrivée du contexte à partir d'une fonction de densité

de probabilité exponentielle

$$f(x|\gamma,\lambda) = \begin{cases} \lambda e^{-\lambda(x-\gamma)} & \text{pour } x \geqslant \gamma \\ 0 & \text{sinon} \end{cases}$$

d'emplacement $\gamma = 3.42$ mm$^3$ et d'échelle $1/\lambda = 66.88$, construite à partir de la distribution des contextes actuellement disponibles (voir la figure 2.2b).

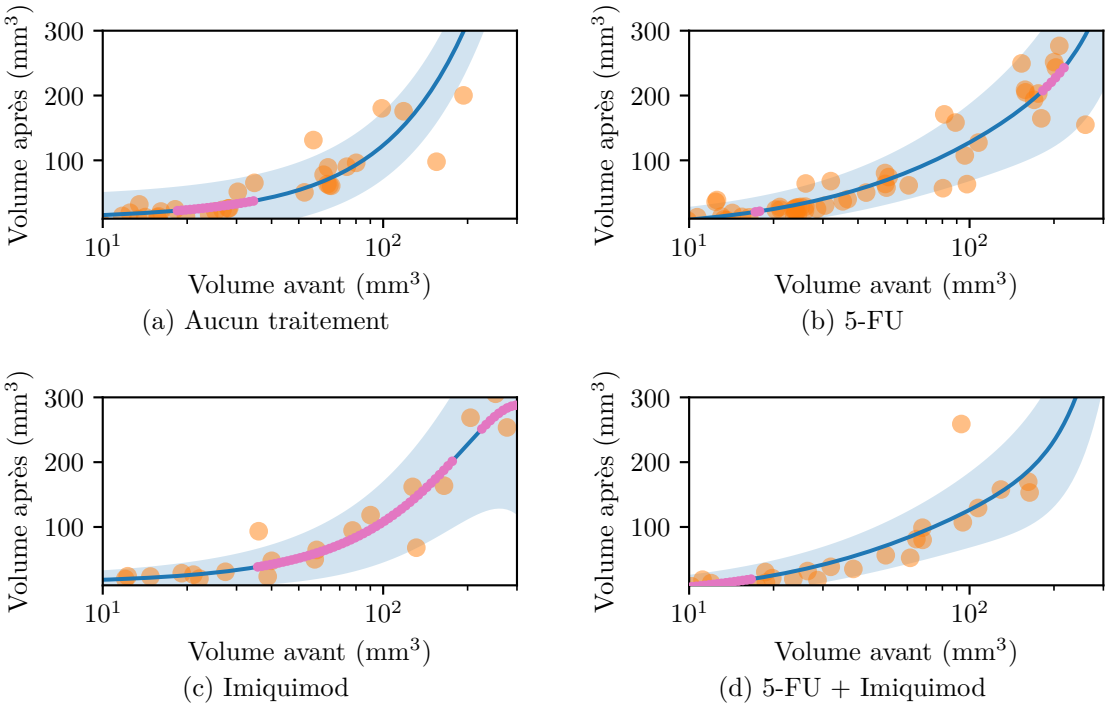
Nous considérons quatre modèles de simulation pour les fonctions de transition moyenne d'un volume initial vers le volume subséquent. Pour chaque traitement $a \in \mathcal{A}$, la fonction $f_a$ est modélisée par une régression linéaire, cubique, quartique et quintique effectuée sur les données obtenues durant la phase de RCT. Dans chacune de ces configurations, le bruit est modélisé comme une fonction linéaire du volume de la tumeur obtenue en effectuant une régression linéaire sur les écarts types entre les données de la phase de RCT et les fonctions $f_a$ modélisées. Les figures 2.3 à 2.6 montrent respectivement les fonctions $f_a$ pour chaque traitement, pour des volumes de tumeur variant de 3 mm$^3$ à 300 mm$^3$. Cela correspond à 99% des contextes présents dans les données obtenues durant la phase RCT (l'autre 1% correspondant à des volumes supérieurs). Les points indiquent les données ayant servi à obtenir chaque modèle. Notons l'échelle logarithmique qui est utilisée pour mettre l'accent des fonctions de transition sur les petits volumes puisqu'ils sont plus fréquents que les tumeurs volumineuses. Un traitement est considéré comme optimal pour un volume de tumeur donné si l'espérance du volume suivant le traitement est inférieure à l'espérance du volume suivant les autres traitements. Les régions roses le long des fonctions indiquent que le traitement en question est optimal pour ce volume de tumeur. On remarque que ces différents modèles impliquent un bruit qui n'est pas constant à travers l'espace des contextes. Cela va à l'encontre de l'hypothèse effectuée par le modèle de régression à noyau classique et nous permettra d'évaluer la robustesse des approches à cette situation non idéale qui peut survenir en pratique.




(a) Aucun traitement (b) 5-FU

(c) Imiquimod (d) 5-FU + Imiquimod

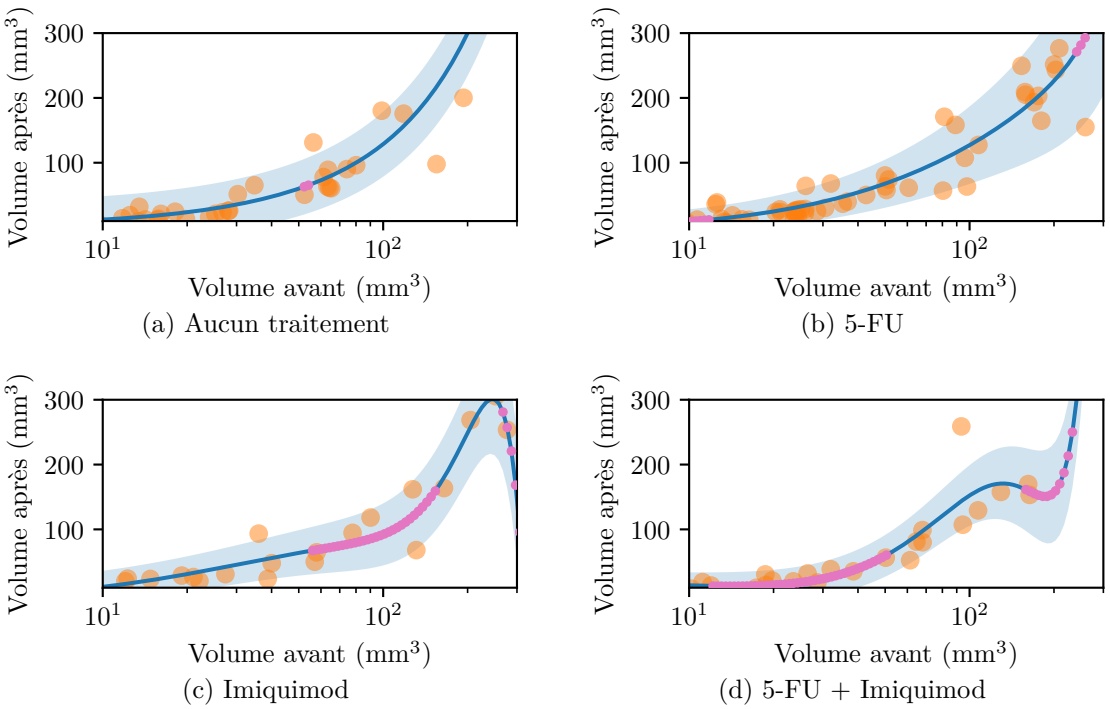
FIGURE 2.3 – Modèles de simulation linéaires (bleu) avec un écart type de bruit, les observations du RCT (orange), ainsi que la région pour laquelle la fonction est optimale (rose).




(a) Aucun traitement (b) 5-FU

(c) Imiquimod (d) 5-FU + Imiquimod

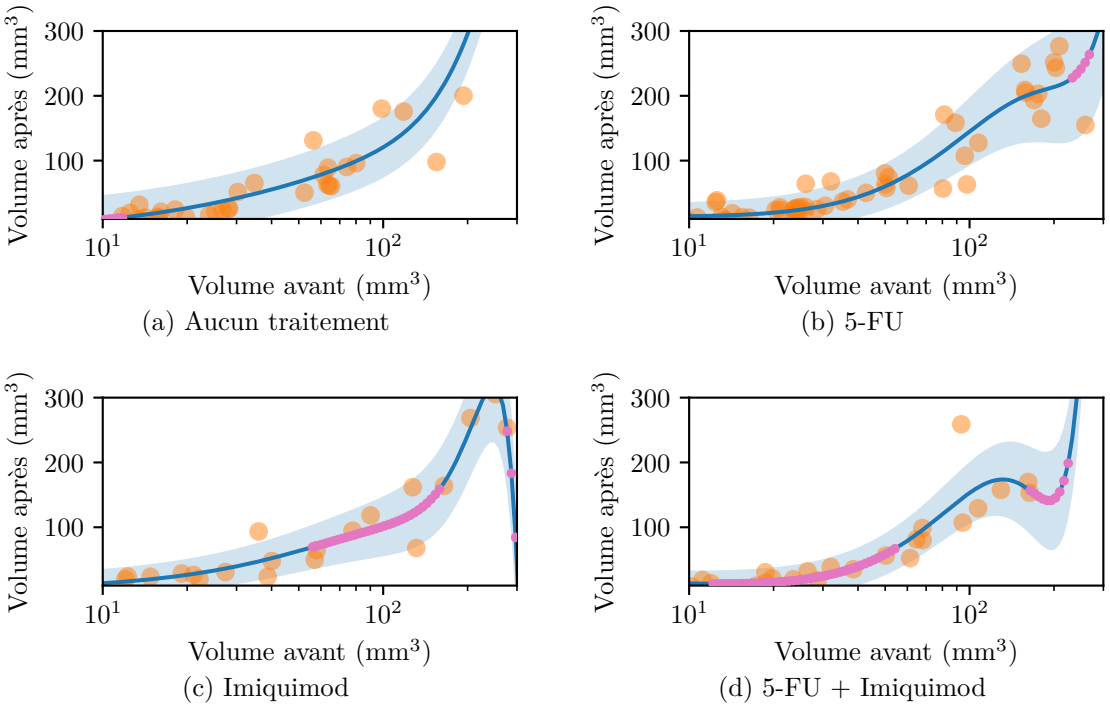
FIGURE 2.4 – Modèles de simulation cubiques (bleu) avec un écart type de bruit, les observations du RCT (orange), ainsi que la région pour laquelle la fonction est optimale (rose).




(a) Aucun traitement (b) 5-FU

(c) Imiquimod (d) 5-FU + Imiquimod

FIGURE 2.5 – Modèles de simulation quartiques (bleu) avec un écart type de bruit, les observations du RCT (orange), ainsi que la région pour laquelle la fonction est optimale (rose).




(a) Aucun traitement (b) 5-FU

(c) Imiquimod (d) 5-FU + Imiquimod

FIGURE 2.6 – Modèles de simulation quintiques (bleu) avec un écart type de bruit, les observations du RCT (orange), ainsi que la région pour laquelle la fonction est optimale (rose).

On considère un GP indépendant par action, dont les hyperparamètres du noyau ainsi que la variance du bruit, $\sigma^2$, sont déterminés par maximisation de la vraisemblance (Rasmussen and Williams, 2006) sur les données provenant du RCT. La performance de GP BESA est comparée à celle de CGP-UCB (ou KernelUCB avec $\lambda = \sigma^2$), qui sélectionne l'action

$$a_t = \arg\max_{a \in \mathcal{A}} \left[ f_{\mathcal{J}_{a,t-1}}(x_t) + \beta_t(\delta) s_{\mathcal{J}_{a,t-1}}(x_t) \right], \tag{2.6}$$

avec $\beta_t(\delta) = \sqrt{2 \ln \frac{4t^2\pi^2}{6\delta}}$ (Krause and Ong, 2011) et $\delta = 0.1$. Rappelons que $\mathcal{J}_{a,t}$ correspond à l'historique des contextes et observations obtenues avec l'action $a$ jusqu'au temps $t$ (inclusivement).

**Remarque 4.** *L'équation 2.6 correspond donc à l'équation 2.2 avec un noyau ne partageant aucune information entre les actions.*

Chaque algorithme est exécuté 100 fois sur un horizon de $T = 1000$ épisodes, simulant des mesures de tumeurs et des recommandations de traitement. Contrairement au vrai problème d'ACT, les observations sont obtenues sans délai. Les contextes échantillonnés sont identiques pour tous les algorithmes pour une exécution donnée. Les algorithmes sont comparés en utilisant le pseudo-regret cumulatif donné par l'équation 2.1. La figure 2.7 montre le pseudo-regret cumulatif, moyenné sur les répétitions, pour les quatre configurations de simulation. Rappelons que le but est de minimiser le regret. La valeur absolue du regret n'est pas pertinente : nous sommes plutôt intéressés par la forme de la courbe de regret. Une allocation aléatoire (uniforme) entre les traitements entraînera typiquement un regret linéaire ; nous visons donc un regret logarithmique (ou sous-linéaire en général). Nous observons que GP BESA se compare favorablement avec CGP-UCB. Ces résultats optimistes ont motivé le déploiement de cette stratégie ACT sur des animaux réels, tel que décrit dans la section suivante.

## 2.6 Expériences animales

Nous utilisons maintenant GP BESA comme stratégie d'allocation de traitements adaptative pour la réalisation d'une expérience en laboratoire avec la procédure expérimentale suivante :

- Des tumeurs cutanées sont induites chez des souris utilisant des agents cancérogènes.
- Une souris commence l'expérience lorsque sa plus grande tumeur atteint 3 mm. Nous désignons cette tumeur comme *tumeur principale*.
- Une souris est sacrifiée lorsque sa plus grande tumeur atteint 10 mm ou si une tumeur interne est observée.
- Deux fois par semaine, la tumeur principale est mesurée et un traitement est recommandé par GP BESA compte tenu de son volume. Les experts en laboratoire utilisent une interface graphique dans laquelle ils saisissent le volume de la tumeur principale et obtiennent le traitement à appliquer.

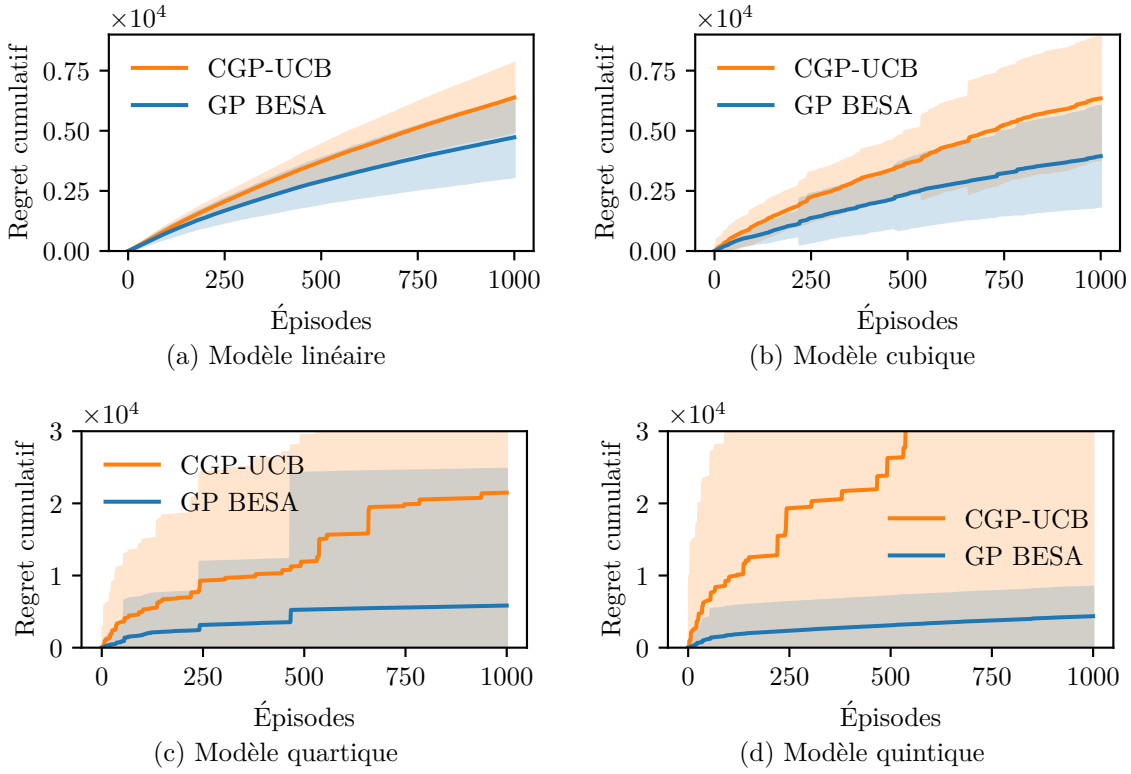



FIGURE 2.7 – Pseudo-regret cumulatif moyen avec un écart type pour les deux heuristiques.

— Chaque traitement alloué est enregistré, avec le volume de la tumeur précédant et suivant le traitement.

La procédure se déroule suivant l'algorithme 6, c'est-à-dire que les données recueillies ne sont pas utilisées immédiatement par GP BESA. Des groupes de souris sont traitées simultanément *au complet* (jusqu'au décès de tous les membres du groupe). Après la terminaison d'un groupe, les données de toutes les souris comprises dans le groupe sont ajoutées à l'historique de GP BESA et utilisées pour la recommandation des traitements pour les souris des groupes subséquents. Dix souris sont traitées au total et regroupées comme suit :

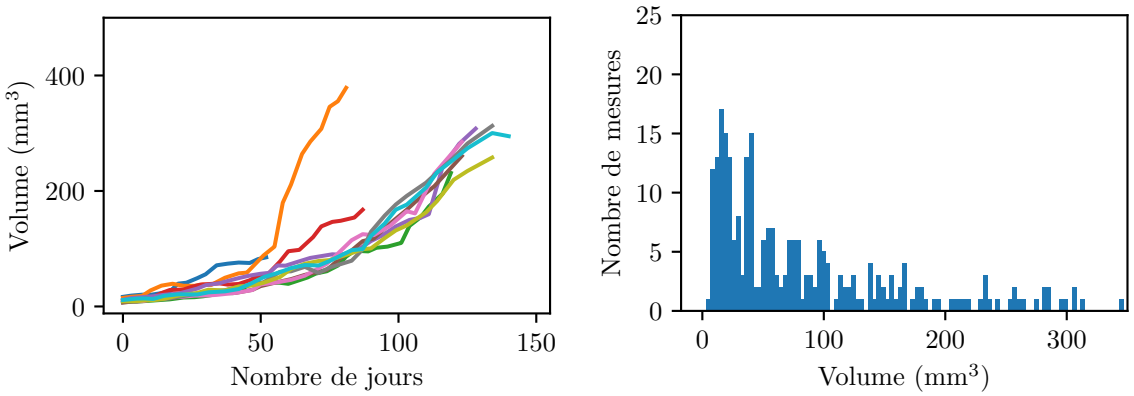
**Groupe A** souris 1 et 2 ;
**Groupe B** souris 3, 4 et 5 ;
**Groupe C** souris 6 et 7 ;
**Groupe D** souris 8, 9 et 10.

L'expérience s'est déroulée sur une durée de 2 ans. L'efficacité de la stratégie obtenue avec GP BESA est comparée aux stratégies de base suivantes.

**Aucun traitement** Un ensemble de données de trois souris est disponible.

**Aléatoire** C'est la politique d'allocation aléatoire utilisée durant la phase RCT, tel que décrit à la section 2.4. Un ensemble de données de cinq souris est disponible.

**5-FU** Un dosage fixe de 5-FU est administré à chaque fois. Cela correspond à toujours choisir




(a) Évolution des tumeurs

(b) Distribution des volumes de tumeur

FIGURE 2.8 – Données acquises durant l'ACT.

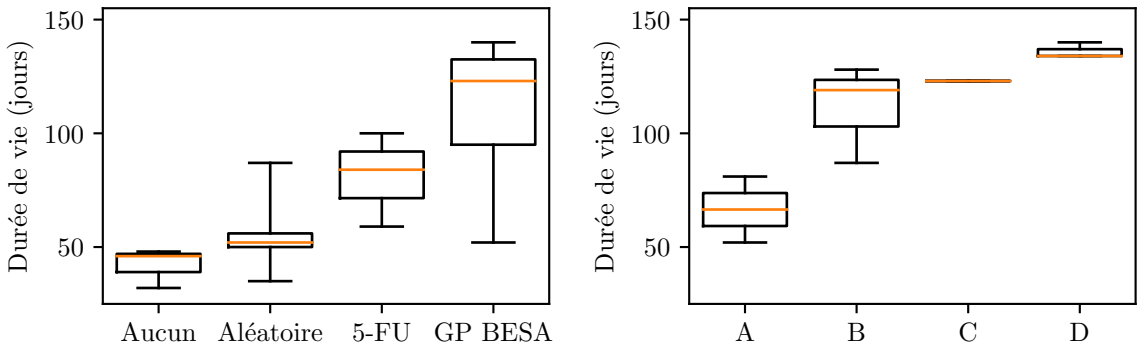
l'action 5-FU comprise dans les choix de traitements offerts à GP BESA. Un ensemble de données de quatre souris est disponible.

Les stratégies sont évaluées en utilisant la durée de vie des animaux comme métrique. Évidemment, on vise à maximiser la durée de vie puisque cela implique une cueillette de données plus importantes. Rappelons que l'on désire maximiser la cueillette d'information de manière à pouvoir éventuellement utiliser ces données pour déterminer des politiques personnalisées.

Les figures 2.8a et 2.8b montrent respectivement la croissance des tumeurs au fil des jours ainsi que la répartition des volumes de tumeurs observés durant la phase d'ACT. En les comparant avec les figures associées pour le RCT (figures 2.2a et 2.2b), on remarque que l'évolution des volumes de tumeurs présente une croissance exponentielle réduite et retardée dans le temps. Cela se traduit par une couverture de l'espace des états plus variée ainsi que des durées de vie plus longue chez les animaux traités avec la procédure d'ACT.

La figure 2.9a montre la répartition des durées de vie des souris pour chaque stratégie d'allocation de traitement de base et GP BESA. Les boîtes indiquent l'écart interquartile (EI), les poignées indiquent l'étendue des données et la barre intermédiaire (orange) représente la médiane. Nous observons d'une part que la stratégie adaptative de GP BESA permet d'augmenter la durée de vie des souris comparativement aux stratégies d'allocation des traitements de base. L'étendue des données obtenues avec GP BESA est cependant beaucoup plus large qu'avec les autres stratégies. Cela s'explique par la figure suivante.

La figure 2.9b montre la répartition des durées de vie des souris pour chaque groupe (entre chaque mise à jour) de GP BESA. Les boîtes indiquent la EI, les poignées indiquent l'étendue des données et la barre intermédiaire (orange) représente la médiane. Nous observons que GP BESA s'améliore après chaque mise à jour, c'est-à-dire après l'intégration des données recueillies sur le groupe de souris précédent. Plus spécifiquement, nous remarquons une augmentation de plus de 50% de longévité entre la médiane de la meilleure stratégie de base



(a) Par stratégie (b) Par groupe de GP BESA

FIGURE 2.9 – Durée de vie médiane des cobayes animaux. Les boîtes couvrent du quartile inférieur au quartile supérieur et les poignées montrent l'étendue des données.

(5-FU) et le dernier groupe de GP BESA. Notons que la différence importante en ce qui concerne la durée de vie pour les souris appartenant aux différents groupes de mise à jour explique l'étendue des données de GP BESA à la figure précédente. Nous remarquons également que les groupes de souris présentent beaucoup moins de variabilité dans leur longévité au fur et à mesure des mises à jour. Cela s'explique peut-être par l'utilisation de stratégies de traitements moins variables avec la convergence de GP BESA. Cet effet devrait cependant être validé sur une cohorte plus importante.

## 2.7 Discussion

Dans ce chapitre, nous avons utilisé des données recueillies initialement dans une phase de RCT pour concevoir une politique d'ACT. Nous avons formulé le problème d'ACT dans le cadre des bandits contextuels, dans lequel nous devions sélectionner parmi quatre traitements en fonction du volume tumoral mesuré. Les rétroactions utilisées pour l'apprentissage correspondent à la transition immédiate de l'état, c'est-à-dire au prochain volume de la tumeur après un traitement. Cette formulation encourage le système à rester dans des états de volume tumoral inférieurs, ce qui permet de recueillir plus d'échantillons en ralentissant la croissance de la tumeur, donc en étirant la durée de vie des sujets animaux. Nous avons utilisé des GPs pour modéliser la fonction de rétroaction de chaque traitement, associant chaque volume de tumeur vers le volume tumoral subséquent, et nous avons proposé GP BESA pour sélectionner le prochain traitement en se basant sur un sous-échantillon d'observations permettant d'obtenir des intervalles de confiance similaires sur tous les traitements.

Les résultats empiriques présentés ici suggèrent que le GP BESA est effectivement en mesure de capturer les caractéristiques très variables de la progression du cancer, même dans le contexte de la thérapie. Il peut donc être un outil utile pour tenter de modéliser la dynamique de la croissance tumorale, ce qui constitue un défi d'actualité (Loizides et al., 2015).

Une meilleure compréhension de l'évolution des tumeurs cancéreuses pourrait aider à la découverte de traitements et favoriser la mise en place de stratégies adaptées à la maladie. Des travaux futurs permettront de confirmer le potentiel d'utilisation des données recueillies pour l'apprentissage hors-ligne de politiques de traitements.

Ces résultats supportent également le déploiement de stratégies d'ACT dans des contextes d'application réels. Bien que les bandits (Thompson, 1933; Robbins, 1952) soient nés d'une application à l'ACT, il n'en demeure pas moins que l'utilisation de leurs algorithmes sur le terrain est encore limitée (Villar et al., 2015). Davantage d'efforts pour offrir des garanties théoriques sous des hypothèses réalisables dans la pratique pourraient contribuer à faciliter l'adoption de ces stratégies. Notamment, des expériences intégrant plus de variables, caractérisant autant de la maladie que les sujets, devraient être effectuées. Des travaux antérieurs (Djolonga et al., 2013; Wang et al., 2016; Li et al., 2016) ont abordé le problème de régression GP sur un espace à grande dimensionnalité. D'autres (Snoek et al., 2015; Springenberg et al., 2016) ont envisagé l'utilisation des réseaux de neurones profonds pour approximer la distribution postérieure. Des extensions de BESA à ce type d'approches pourraient permettre d'étendre les bandits contextuels à des espaces de contextes de dimension trop importante pour les algorithmes actuels.

# Chapitre 3

# Les bandits structurés

De nombreuses applications nécessitent de résoudre un problème d'optimisation d'une fonction inconnue, bruitée, définie sur un espace potentiellement vaste et coûteuse à évaluer. La solution consiste généralement à s'appuyer sur une forme de régression pour obtenir un modèle de la fonction à optimiser. Ce dernier est ensuite utilisé pour la prise de décision et la sélection du prochain point à observer. Dans certaines situations, on dispose d'un certain nombre (limité) d'essais pour échantillonner la fonction à différents points, après quoi le but est d'identifier avec le plus de précision possible le point optimisant la fonction. Ce type de problème est dit *pure exploration*, puisque la phase d'échantillonnage ne sert qu'à explorer ladite fonction dans le but de maximiser sa connaissance.

Dans ce chapitre, nous abordons le problème d'optimisation dans lequel les essais ont un *impact*, c'est-à-dire que la fonction doit être optimisée simultanément à son échantillonnage. Par exemple, lors de l'apprentissage du dosage optimal d'un traitement pour des patients atteints d'une maladie donnée, les essais contribuent simultanément à parfaire l'estimation de la fonction d'efficacité du dosage et à traiter les patients. Il est donc nécessaire d'effectuer un compromis entre des essais potentiellement optimaux, soit exploiter la connaissance actuelle de la fonction, et des essais permettant d'améliorer la connaissance de la fonction, soit explorer des régions de l'espace d'entrée qui sont méconnues. Nous formulons ce problème comme des bandits aux actions structurées, dans lequel l'espace des actions correspond à l'espace d'entrée d'une fonction à optimiser. Plus spécifiquement, nous nous intéressons à la situation dans laquelle la fonction appartient à un *reproducing kernel Hilbert space* (RKHS). Aussi connu sous le nom de bandits à noyau (*kernelized bandits*), ce problème repose sur la régression à noyau pour l'estimation de la fonction à optimiser.

Il est connu que le paramètre de régularisation dans le processus de régression à noyau dépende de la variance du bruit sur la fonction observée. En pratique, quand la variance du bruit est inconnue, la stratégie naturelle consiste à l'estimer. Donc, naturellement, la régularisation devrait pouvoir s'adapter à cette estimation empirique du bruit. Malheureusement, les inter-

valles de confiance existants (Srinivas et al., 2010; Valko et al., 2013; Wang and de Freitas, 2014) sur la régression à noyau requièrent que le paramètre de régularisation soit fixe. Nous nous attaquons à cette question en introduisant d'abord un résultat de concentration pour la régression à noyau avec un paramètre de régularisation variable. Nous proposons ensuite une technique basée sur des estimateurs empiriques de la variance permettant d'ajuster le paramètre de régularisation de manière adaptative en fonction des observations passées.

Nous proposons par la suite une variante de Thompson *sampling* (TS) basée pour les bandits structurés en RKHS : Kernel TS. En utilisant les résultats de concentration introduits précédemment, nous fournissons une analyse théorique de Kernel TS pour l'optimisation de fonctions de bruit inconnu à régularisation adaptative. Dans la même lignée que des travaux antérieurs (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016), l'algorithme de TS résultant implique un gonflement artificiel de la covariance postérieure, contrairement à l'échantillonnage direct de la distribution postérieure, typique dans le TS classique.

Des expériences empiriques sont finalement réalisées pour évaluer l'impact d'une régularisation adaptative et des résultats de concentration introduits avec différentes approches de bandits. La convergence de Kernel TS, possédant des garanties de convergence, est ensuite comparée à une version alternative sans gonflement de la covariance, mais pour laquelle aucune garantie n'est offerte.

## 3.1 Formulation du problème

Un problème de bandits structurés est décrit par un ensemble (compact) d'actions $\mathcal{X} \subset \mathbb{R}^d$ et une fonction cible à optimiser (inconnue) $f_\star : \mathcal{X} \mapsto \mathbb{R}$. Nous supposons que la fonction cible $f_\star$ appartient au RKHS d'un noyau $k$ (voir la section 1.2.3). À chaque épisode $t \in \mathbb{N}_{>0}$, un agent choisit d'échantillonner la fonction au point $x_t$ et obtient une observation $y_t$ perturbée par le bruit $\xi_t$, telle que $y_t := f_\star(x_t) + \xi_t$.

Un algorithme de bandits structurés est une méthode (possiblement randomisée) pour sélectionner le prochain point à échantillonner étant donné l'historique des choix antérieurs et des observations obtenues, $\mathcal{H}_t := \{x_s, y_s\}_{s=1}^{t-1}$. Soit l'action optimale $\star$ pour une fonction cible $f_\star$ donnée, c'est-à-dire $\star := \arg\max_{x \in \mathcal{X}} f_\star(x)$. Le but de l'algorithme est de minimiser le pseudo-regret cumulatif :

$$\mathfrak{R}(T) \overset{\text{def}}{=} \sum_{t=1}^{T} [f_\star(\star) - f_\star(x_t)]. \tag{3.1}$$

Cette quantité mesure la performance d'un algorithme comparée à celle d'un oracle qui connaîtrait la fonction cible $f_\star$, donc qui connaîtrait $\star$.

**Remarque 5.** *$f_\star(x)$ correspond à l'observation attendue au point $x$, tel que $f_\star(x) = \mu_x$ dans la formulation du bandit classique. L'équation 3.1 correspond donc directement à la définition 3.*

**Hypothèses** Pour la suite, nous supposons que la fonction cible $f_\star$ appartient au RKHS $\mathcal{K}$ et que le noyau est borné à 1, c'est-à-dire $k(x,x) \leqslant 1$ pour tout $x \in \mathcal{X}$.

## 3.2 Littérature

Aussi connu sous le nom des bandits aux actions continues (*continuum-armed* bandits), bandits à $\mathcal{X}$-actions ($\mathcal{X}$-*armed bandits*), ce problème a été abordé précédemment sous l'hypothèse d'une fonction cible Lipschitz ou Hölder, localement ou globalement (Agrawal, 1995; Kleinberg, 2004; Kleinberg et al., 2008; Bubeck et al., 2011a,b; Magureanu et al., 2014). Les bandits structurés ont également été abordés précédemment sous l'hypothèse d'une structure linéaire. À cet effet, des algorithmes basés sur *upper confidence bound* (UCB) (Abbasi-Yadkori et al., 2011; Chu et al., 2011) et sur TS (Agrawal and Goyal, 2013; Abeille and Lazaric, 2016) ont été proposées.

L'approche GP-UCB (Srinivas et al., 2010) généralise UCB du cadre linéaire au cas RKHS à l'aide de *Gaussian processs* (GPs). Plus spécifiquement, GP-UCB sélectionne l'action

$$x_t = \arg\max_{x \in \mathcal{X}} \Big[ f_{t-1}(x) + B_t(\delta) s_{t-1}(x) \Big], \tag{3.2}$$

où $f_{t-1}(\cdot)$ et $s_{t-1}(\cdot)$ représentent respectivement la moyenne et l'écart-type a posteriori d'un GP (voir les équations 1.27 et 1.30), et le second terme correspond à un intervalle de confiance sur l'estimateur $f_{t-1}(x)$ tel que

$$\mathbb{P}\Big[ |f_{t-1}(x) - f_\star(x)| > B_t(\delta) s_{t-1}(x) \Big] \leqslant \delta$$

pour tout $\delta \in [0,1]$. Les auteurs fournissent des analyses théoriques sous l'hypothèse que la fonction à optimiser est échantillonnée d'un GP et sous l'hypothèse que la fonction à optimiser appartient au RKHS $\mathcal{K}$. Cependant, ces analyses se limitent à l'hypothèse de bruit borné. L'approche KernelUCB (Valko et al., 2013) généralise également UCB du cadre linéaire au cas RKHS en utilisant la régression à noyau avec $\lambda$ général. KernelUCB sélectionne l'action

$$x_t = \arg\max_{x \in \mathcal{X}} \left[ f_{\lambda,t-1}(x) + B_{\lambda,t}(\delta) \sqrt{\frac{k_{\lambda,t-1}(x,x)}{\lambda}} \right], \tag{3.3}$$

où $f_{\lambda,t-1}(\cdot)$ et $k_{\lambda,t-1}(\cdot)$ sont respectivement données par les équations 1.32 et 1.33, et le second terme correspond à un intervalle de confiance sur l'estimateur $f_{\lambda,t-1}(x)$ tel que

$$\mathbb{P}\left[ |f_{\lambda,t-1}(x) - f_\star(x)| > B_{\lambda,t}(\delta) \sqrt{\frac{k_{\lambda,t-1}(x,x)}{\lambda}} \right] \leqslant \delta$$

pour tout $\delta \in [0,1]$. L'analyse de KernelUCB permet d'obtenir des bornes sur le regret plus serrées que CGP-UCB dans le cas RKHS mais elle se limite à l'hypothèse d'observations bornées. Cela implique non seulement que le bruit est borné, mais également que la fonction cible $f_\star$ est bornée.

## 3.3 Régression à noyau et régularisation adaptative

Considérons le résultat de concentration suivant pour la régression à noyau, sous l'hypothèse que le bruit $\xi_t$ est $\sigma$-sous-gaussien [1] et indépendant de $x_t$ tel que

$$\ln \mathbb{E}[\exp(\gamma \xi_t) | \mathcal{H}_t] \leqslant \frac{\gamma^2 \sigma^2}{2} \quad \forall t \in \mathbb{N}_{>0}, \forall \gamma \in \mathbb{R},$$

pour $\sigma \geqslant 0$. Pour la suite, rappelons que $\boldsymbol{I}_n$ dénote une matrice identité de taille $n \times n$.

**Lemme 2** (Régression à noyau (Maillard, 2016))**.** *Sous l'hypothèse de bruit $\sigma$-sous-gaussien, pour un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$, la moyenne et la variance postérieure suite aux observations $\mathbf{y}_t \in \mathbb{R}^{t \times 1}$ sont données par*

$$f_{\lambda,t}(x) = \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda \boldsymbol{I}_t)^{-1} \mathbf{y}_t$$
$$s^2_{\lambda,t}(x) = \frac{\sigma^2}{\lambda} k_{\lambda,t}(x,x) \quad \textit{avec} \quad k_{\lambda,t}(x,x') = k(x,x') - \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda \boldsymbol{I}_t)^{-1} \mathbf{k}_t(x'),$$

*où $\mathbf{k}_t(x) = (k(x,x_s))_{s \leqslant t}$ et $\mathbf{K}_t = (K(x_s, x_{s'}))_{s,s' \leqslant t}$. Alors l'inégalité suivante est vraie simultanément pour tout $x \in \mathcal{X}$ et pour tout $t \geqslant 0$ :*

$$|f_\star(x) - f_{\lambda,t}(x)| \leqslant \sqrt{\frac{k_{\lambda,t}(x,x)}{\lambda}} \Big[ \sigma \sqrt{2 \ln(1/\delta) + 2\gamma_t(\lambda)} + \sqrt{\lambda} \|f_\star\|_{\mathcal{K}} \Big],$$

*avec probabilité supérieure à $1-\delta$, pour tout $\delta \in [0,1]$. Rappelons que $\gamma_t(\lambda) = \frac{1}{2} \sum_{s=1}^t \ln\left(1 + \frac{1}{\lambda} k_{\lambda,s-1}(x_s, x_s)\right)$ dénote le gain d'information (voir la section 1.2.5).*

Le cas où $\lambda = \lambda_\star \overset{\text{def}}{=} \sigma^2 / \|f_\star\|^2_{\mathcal{K}}$ est particulièrement intéressant puisque nous obtenons alors

$$f_{\lambda_\star,t}(x) = \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda_\star \boldsymbol{I}_t)^{-1} \mathbf{y}_t$$
$$s^2_{\lambda_\star,t}(x) = \|f_\star\|^2_{\mathcal{K}} k_{\lambda_\star,t}(x,x) \quad \text{avec} \quad k_{\lambda_\star,t}(x,x') = k(x,x') - \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda_\star \boldsymbol{I}_t)^{-1} \mathbf{k}_t(x')$$

ainsi que

$$|f_\star(x) - f_{\lambda_\star,t}(x)| \leqslant \|f_\star\|_{\mathcal{K}} \sqrt{k_{\lambda_\star,t}(x,x)} \Big[ \sqrt{2 \ln(1/\delta) + 2\gamma_t(\lambda_\star)} + 1 \Big].$$

Des résultats numériques présentés à la section 3.6.1 illustrent la pertinence de ce choix de régularisation en comparant son impact sur les intervalles de confiance avec l'intuition bayésienne classique ($\lambda = \sigma^2$). Rappelons que la norme d'une fonction est liée à sa complexité, c'est-à-dire son irrégularité (voir la section 1.2.3). Rappelons également que la complexité du modèle de régression a posteriori est liée à la régularisation (voir la section 1.2.5). Le fait de considérer la norme de la fonction dans la régularisation permet donc au modèle de régression de s'adapter non seulement au bruit, mais également à la complexité de la fonction.

1. La sous-gaussianité implique une certaine convergence des queues d'une distribution.

Cependant, en pratique, $\|f_\star\|_{\mathcal{K}}^2$ et $\sigma^2$ ne sont généralement pas connus *exactement*. Nous supposerons plus tard qu'une borne supérieure est donnée sur $\|f_\star\|_{\mathcal{K}}$. Nous voulons alors maintenir une estimation de $\sigma^2$ à chaque épisode $t$, de manière à pouvoir ajuster $\lambda$ pour l'acquisition du prochain point $x_{t+1}$. L'utilisation d'une séquence de paramètres de régularisation $(\lambda_t)_{t\geqslant 1}$ ajustée de manière adaptative en se basant sur les observations antérieures requiert de modifier le résultat de concentration précédent (lemme 2), qui est valide seulement pour une régularisation déterministe. Nous obtenons donc le résultat plus général suivant.

**Théorème 1** (Régression à noyau avec régularisation adaptative)**.** *Considérons les mêmes hypothèses que le lemme 2. Soit la séquence* prédictible *de paramètres positifs* $\boldsymbol{\lambda} = (\lambda_t)_{t\geqslant 1}$*, c'est-à-dire que* $\lambda_t$ *est déterminée par le passé* $\mathcal{H}_{t-1}$*. Supposons que pour chaque* $t$*,* $\lambda_t \geqslant \lambda_\star$ *est vrai pour une constante positive* $\lambda_\star$*. La moyenne et la variance postérieure suite aux observations* $\mathbf{y}_t \in \mathbb{R}^{t\times 1}$ *sont données par*

$$f_{\boldsymbol{\lambda},t}(x) = \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda_{t+1}\boldsymbol{I}_t)^{-1}\mathbf{y}_t$$
$$s^2_{\boldsymbol{\lambda},t}(x) = \frac{\sigma^2}{\lambda_{t+1}} k_{\lambda_{t+1},t}(x,x) \quad \textit{avec} \quad k_{\lambda,t}(x,x') = k(x,x') - \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda \boldsymbol{I}_t)^{-1}\mathbf{k}_t(x'),$$

*où* $\mathbf{k}_t(x) = (k(x,x_s))_{s\leqslant t}$ *et* $\mathbf{K}_t = (K(x_s,x_{s'}))_{s,s'\leqslant t}$*. Alors l'inégalité suivante est vraie simultanément pour tout* $x \in \mathcal{X}$ *et pour tout* $t \geqslant 0$ *:*

$$|f_\star(x) - f_{\boldsymbol{\lambda},t}(x)| \leqslant \sqrt{\frac{k_{\lambda_{t+1},t}(x,x)}{\lambda_{t+1}}}\Big[\sigma\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda_\star)} + \sqrt{\lambda_{t+1}}\|f_\star\|_{\mathcal{K}}\Big],$$

*avec probabilité supérieure à* $1-\delta$*, pour tout* $\delta \in [0,1]$*. Rappelons que* $\gamma_t(\lambda) = \frac{1}{2}\sum_{s=1}^t \ln\left(1 + \frac{1}{\lambda}k_{\lambda,s-1}(x_s,x_s)\right)$ *dénote le gain d'information (voir la section 1.2.5).*

*Démonstration.* La preuve repose sur une extension directe de l'analyse existante pour une régularisation $\lambda$ déterministe (Maillard, 2016). Soit $\theta_\star$ le vecteur de paramètre correspondant à la fonction $f_\star \in \mathcal{K}$, tel que $f_\star(x) = \theta_\star^\top \varphi(x)$. Soient la matrice des caractéristiques $\boldsymbol{\Phi}_t = (\varphi(x_s))_{s\leqslant t}$ de dimension $t \times \infty$, la matrice bi-infinie $\mathbf{V}_{\lambda,t} = \boldsymbol{I}_\infty + \frac{1}{\lambda}\boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t$, ainsi que le vecteur de bruits $E_t = (\xi_1,\dots,\xi_t)$. Suivant Maillard (2016), nous avons

$$\begin{aligned}
f_{\lambda,t}(x) &= \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda \boldsymbol{I}_t)^{-1}\mathbf{y}_t \\
&= \varphi(x)^\top \boldsymbol{\Phi}_t^\top (\boldsymbol{\Phi}_t\boldsymbol{\Phi}_t^\top + \lambda \boldsymbol{I}_t)^{-1}\mathbf{y}_t \\
&= \varphi(x)^\top \boldsymbol{\Phi}_t^\top \left(\frac{\boldsymbol{I}_t}{\lambda} - \frac{1}{\lambda}\boldsymbol{\Phi}_t(\lambda \boldsymbol{I}_\infty + \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t^\top\right)\mathbf{y}_t \\
&= \varphi(x)^\top (\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t + \lambda \boldsymbol{I}_\infty)^{-1}\boldsymbol{\Phi}_t^\top(\boldsymbol{\Phi}_t\theta_\star + E_t),
\end{aligned}$$

où la troisième égalité est obtenue avec la formule de Sherman-Morrison-Woodbury (lemme 15, annexe A). Ainsi nous obtenons

$$f_{\lambda,t}(x) - f_\star(x) = \frac{1}{\lambda}\varphi(x)^\top \mathbf{V}_{\lambda,t}^{-1}(\boldsymbol{\Phi}_t^\top E_t - \lambda\theta_\star).$$

L'inégalité d'Hölder et la norme pondérée [2] permettent ensuite d'obtenir

$$|f_{\lambda,t}(x) - f_\star(x)| \leqslant \frac{1}{\sqrt{\lambda}}\|\varphi(x)\|_{\mathbf{V}_{\lambda,t}^{-1}}\left(\frac{1}{\sqrt{\lambda}}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda,t}^{-1}} + \sqrt{\lambda}\|\theta_\star\|_{\mathbf{V}_{\lambda,t}^{-1}}\right),$$

où $\|\varphi(x)\|^2_{\mathbf{V}_{\lambda,t}^{-1}} = k_{\lambda,t}(x,x)$ avec Sherman-Morrison-Woodbury, et $\|\theta_\star\|_{V_{\lambda,t}^{-1}} \leqslant \|\theta_\star\|$. Ainsi

$$|f_{\lambda,t}(x) - f_\star(x)| \leqslant \underbrace{\sqrt{\frac{k_{\lambda,t}(x,x)}{\lambda}}}_{\text{(A)}}\Big[\underbrace{\frac{1}{\lambda}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda,t}^{-1}}}_{\text{(B)}} + \underbrace{\sqrt{\lambda}\|\theta_\star\|}_{\text{(C)}}\Big].$$

Les termes (A) et (C) s'adaptent naturellement à n'importe quelle régularisation $\lambda$. Pour le terme (B), nous considérons le lemme suivant, défini pour un paramètre $\lambda$ déterministe (indépendant des observations passées).

**Lemme 3** (Contrôle d'une quantité auto-normalisée (Maillard, 2016))**.** *Sous l'hypothèse d'une séquence $E_t = \{\xi_s\}_{t=0}^\infty$ de bruit $\sigma$-sous-gaussien, pour un paramètre de régularisation $\lambda \in \mathbb{R}_{>0}$ déterministe. Soient les matrices $\boldsymbol{\Phi}_t = (\varphi(x_s)^\top)_{s\leqslant t}$ de dimension $t\times\infty$ et $\mathbf{V}_{\lambda,t} = \boldsymbol{I}_\infty + \frac{1}{\lambda}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t$ de dimension bi-infinie. Alors*

$$\mathbb{P}\left[\exists t : \|\boldsymbol{\Phi}_t^\top E_t\|^2_{\mathbf{V}_{\lambda,t}^{-1}} > 2\sigma^2\lambda\ln(1/\delta) + \sigma^2\lambda\sum_{s=1}^t \ln\left(1 + \frac{1}{\lambda}k_{\lambda,s-1}(x_s,x_s)\right)\right] \leqslant \delta.$$

Sous l'hypothèse qu'il existe une régularisation minimale $\lambda_\star$ telle que $\lambda \geqslant \lambda_\star$, nous avons alors

$$\begin{aligned}
\frac{1}{\lambda}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda,t}^{-1}} &= E_t^\top\frac{\boldsymbol{\Phi}_t^\top}{\lambda}(\boldsymbol{I}_\infty + \frac{1}{\lambda}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t E_t \\
&= E_t^\top\boldsymbol{\Phi}_t^\top(\lambda\boldsymbol{I}_\infty + \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t E_t \\
&\leqslant E_t^\top\boldsymbol{\Phi}_t^\top(\lambda_\star\boldsymbol{I}_\infty + \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t E_t \\
&= \frac{1}{\lambda_\star}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda_\star,t}^{-1}}
\end{aligned}$$

Ainsi, $\frac{1}{\sqrt{\lambda}}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda,t}^{-1}} \leqslant \frac{1}{\sqrt{\lambda_\star}}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda_\star,t}^{-1}}$. Comme $\lambda_\star$ est une quantité déterministe, nous pouvons utiliser le lemme 3 pour obtenir :

$$\mathbb{P}\left[\exists t : \frac{1}{\sqrt{\lambda_\star}}\|\boldsymbol{\Phi}_t^\top E_t\|_{\mathbf{V}_{\lambda_\star,t}^{-1}} > \sigma\sqrt{2\ln(1/\delta) + \sum_{s=1}^t \ln\left(1 + \frac{1}{\lambda_\star}k_{\lambda_\star,s-1}(x_s,x_s)\right)}\right] \leqslant \delta.$$

□

**Remarque 6.** *Ce résultat est valide pour pratiquement n'importe quelle procédure adaptative d'ajustement de la régularisation, la seule contrainte étant l'existance une borne inférieure $\lambda_\star$ sur la régularisation.*

2. Une matrice symmétrique positive définie $W$ induit la norme pondérée $\|x\|_W = \sqrt{x^\top Wx} = \|W^{1/2}x\|_2$.

**Remarque 7.** *La condition requise $\lambda_t \geqslant \lambda_\star$ sera naturellement satisfaite par le choix de régularisation que nous allons considérer juste ici.*

Supposons maintenant que nous possédions une borne supérieure $C$ sur la norme $f_\star$, ainsi qu'une borne supérieure la variance du bruit, $\tilde{\sigma}_{+,t}$, construite à partir des observations recueillies jusqu'au temps $t$ (inclusivement). Les résultats précédents suggèrent de définir la séquence $(\lambda_t)_{t\geqslant 1}$ avec

$$\lambda_t = \sigma_{+,t-1}^2/C^2 \quad \text{avec} \quad \sigma_{+,t} = \min\{\tilde{\sigma}_{+,t}, \sigma_{+,t-1}\} \quad \text{et} \quad \sigma_{+,0} = \sigma_+, \tag{3.4}$$

où $\sigma_+ \geqslant \sigma$ est une borne supérieure initiale (et lâche) sur $\sigma$. Ainsi, $\lambda_t$ est déterminée par les observations passées et satisfait $\lambda_t \geqslant \lambda_\star$ avec grande probabilité pour $\lambda_\star = \sigma^2/C^2$. Soit l'estimateur empirique (légèrement biaisé) de la variance

$$\widehat{\sigma}_{\lambda,t} = \frac{1}{t}\sum_{s=1}^{t}\Big(y_s - f_{\lambda,t}(x_s)\Big)^2, \tag{3.5}$$

pour une régularisation $\lambda$. Sous l'hypothèse d'un bruit $\sigma$-sous-gaussien et $\sigma$-sous-gaussien de second ordre [3] tel que

$$\ln \mathbb{E}[\exp(\gamma\xi_t^2)|\mathcal{H}_t] \leqslant -\frac{1}{2}\ln(1-2\gamma^2\sigma^2), \quad \forall t \in \mathbb{N}_{>0}, \forall \gamma < \frac{1}{2\sigma^2},$$

pour $\sigma \geqslant 0$, nous considérons les estimateurs résumés dans le lemme suivant.

**Remarque 8.** *Pour éviter les formalités, on peut supposer que $\xi_t \sim \mathcal{N}(0,\sigma^2)$.*

**Lemme 4** (Estimation empirique de la variance par régression à noyau (Maillard, 2016))**.** *Soit une séquence prédictible positive $\boldsymbol{\lambda} = (\lambda_t)_{t\geqslant 1}$, telle que $\lambda_t \geqslant \lambda_\star$ est vrai pour tout $t$. Nous définissons les quantités*

$$C_t(\delta) = \ln(e/\delta)\big[1+\ln(\pi^2\ln(t)/6)/\ln(1/\delta)\big], \quad D_{\lambda,t}(\delta) = 2\ln(1/\delta) + 2\gamma_t(\lambda)$$
$$\text{et} \quad \alpha = \max\Big(1 - \sqrt{\frac{C_t(\delta')}{t}} - \sqrt{\frac{C_t(\delta')+2D_{\lambda_\star,t}(\delta')}{t}}, 0\Big).$$

*Nous considérons ensuite les bornes suivantes sur la variance du bruit, définies différemment, tout dépendant si une borne supérieure $\sigma_+ \geqslant \sigma$ est connue (cas 1) ou non (cas 2).*

$$\sigma_{+,t}(\lambda,\lambda_\star) = \begin{cases} \widehat{\sigma}_{\lambda,t} + \sigma_+\Big(\sqrt{\frac{C_t(\delta')}{t}} + \sqrt{\frac{C_t(\delta')+2D_{\lambda_\star,t}(\delta')}{t}}\Big) + \sqrt{\frac{2\sigma_+\|f_\star\|_{\mathcal{K}}\sqrt{\lambda D_{\lambda_\star,t}(\delta')}}{t}} & \textit{(cas 1)} \\ \frac{1}{\alpha^2}\left(\sqrt{\widehat{\sigma}_{\lambda,t}\alpha + \frac{\|f_\star\|_{\mathcal{K}}\sqrt{\lambda D_{\lambda_\star,t}(\delta')}}{2t}} + \sqrt{\frac{\|f_\star\|_{\mathcal{K}}\sqrt{\lambda D_{\lambda_\star,t}(\delta')}}{2t}}\right)^2 & \textit{(cas 2)} \end{cases}$$

$$\sigma_{-,t}(\lambda) = \begin{cases} \widehat{\sigma}_{\lambda,t} - \sigma_+\sqrt{\frac{2C_t(\delta')}{t}} - \|f_\star\|_{\mathcal{K}}\sqrt{\frac{\lambda}{t}\Big(1 - \frac{1}{\max_{t'\leqslant t}(1+\frac{1}{\lambda}k_{\lambda,t'-1}(x_{t'},x_{t'}))}\Big)} & \textit{(cas 1)} \\ \Big[\widehat{\sigma}_{\lambda,t} - \|f_\star\|_{\mathcal{K}}\sqrt{\frac{\lambda}{t}\Big(1 - \frac{1}{\max_{t'\leqslant t}(1+\frac{1}{\lambda}k_{\lambda,t'-1}(x_{t'},x_{t'}))}\Big)}\Big]\Big(1+\sqrt{\frac{2C_t(\delta')}{t}}\Big)^{-1} & \textit{(cas 2).} \end{cases}$$

3. On reconnaît à droite la fonction génératrice des moments de la distribution $\chi^2$ de degré 1. Cette hypothèse est naturellement valide pour un bruit gaussien.

*Sous l'hypothèse de bruit $\sigma$-sous-gaussien de second ordre et d'un modèle de régression à noyau prédictible, alors*

$$\sigma_{-,t}(\lambda_t) \leqslant \sigma \leqslant \sigma_{+,t}(\lambda_t, \lambda_\star)$$

*simultanément pour tout $t \geqslant 0$ avec probabilité supérieure à $1 - 3\delta$.*

**Remarque 9.** *Ces quantités demeureront des bornes supérieures et inférieures en remplaçant la norme de $f_\star$ par une borne supérieure $C \geqslant \|f_\star\|_{\mathcal{K}}$.*

**Remarque 10.** *Le terme $\|f_\star\|_{\mathcal{K}}$ apparaît systématiquement accompagné du facteur $\sqrt{\lambda}$. Cela suggère de choisir $\lambda$ proportionnel à $1/\|f_\star\|_{\mathcal{K}}^2$, appuyant l'idée de viser $\lambda_\star = \sigma^2/C^2$, où $C$ est une borne supérieure sur la norme de $f_\star$.*

**Remarque 11.** *Sachant que l'intervalle de confiance d'un estimateur empirique de la variance $\widehat{\sigma}_{\lambda,t}$ est d'ordre $\mathcal{O}(1/\sqrt{t})$ (Maurer and Pontil, 2009), nous avons $\sigma_{-,t}(\lambda) = \sigma - \mathcal{O}(1/\sqrt{t})$ et $\sigma_{+,t}(\lambda, \lambda_\star) = \sigma + \mathcal{O}(1/\sqrt{t})$.*

On observe que l'estimation de la borne supérieure $\sigma_{+,t}(\lambda, \lambda_\star)$ dépend de $\lambda_\star$, soit une borne inférieure sur $\lambda_t$. Nous considérons alors une borne inférieure de $\sigma$ construite à partir des observations recueillies jusqu'au temps $t$ (inclusivement),

$$\sigma_{-,t} = \max\{\tilde{\sigma}_{-,t}, \sigma_{-,t-1}\} \quad \text{avec} \quad \sigma_{-,0} = \sigma_-, \tag{3.6}$$

où $0 \leqslant \sigma_- \leqslant \sigma$ est une borne inférieure initiale (et lâche) sur $\sigma$. Une façon de procéder consiste alors à calculer une estimation $\tilde{\sigma}_{-,t} = \sigma_{-,t}(\lambda)$, laquelle peut ensuite être utilisée pour calculer la quantité $\lambda_- \leqslant \lambda_\star$, puis calculer l'estimation $\tilde{\sigma}_{+,t} = \sigma_{+,t}(\lambda, \lambda_-) \geqslant \sigma_{+,t}(\lambda, \lambda_\star)$. Il est alors possible de construire une séquence prédictible $\boldsymbol{\lambda}$ telle que décrite par l'équation 3.4.

Suivant le lemme 4, nous avons $\sigma_{-,t} \leqslant \sigma$ tel que $\lambda_- \leqslant \lambda_\star$ et $\sigma_{+,t} \geqslant \sigma$, donc $\lambda_t \geqslant \lambda_\star$ simultanément pour tout $t \geqslant 0$ avec grande probabilité. En utilisant une *union bound* (inégalité de Boole) pour remplacer la variance $\sigma^2$ par son estimation $\sigma_{+,t}$ dans le théorème 1, nous obtenons des bornes de confiance entièrement calculables dans un contexte où le paramètre de régularisation est ajusté de manière adaptative et où le bruit sur la fonction observée est inconnu. Le corollaire suivant résume ce résultat.

**Corollaire 1** (Régression à noyau avec régularisation adaptative empirique)**.** *Rappelons que $C \geqslant \|f_\star\|_{\mathcal{K}}$. Pour chaque $t \geqslant 1$, définissons la borne inférieure sur le bruit,*

$$\sigma_{-,t} = \max\{\sigma_{-,t}(\lambda_{t-1}), \sigma_{-,t-1}\}, \quad \textit{avec} \quad \sigma_{-,0} = \sigma_- \leqslant \sigma,$$

*ainsi que la borne inférieure correspondante sur $\lambda_\star$, $\lambda_- = \sigma_{-,t}^2/C^2$. Puis, pour chaque $t \geqslant 1$, définissons la borne supérieure sur le bruit,*

$$\sigma_{+,t} = \min\{\sigma_{+,t}(\lambda_{t-1}, \lambda_-), \sigma_{+,t-1}\} \quad \textit{avec} \quad \sigma_{+,0} = \sigma_+ \geqslant \sigma.$$

*Nous définission alors le paramètre de régularisation du modèle de régression utilisé pour l'acquisition de l'observation $t+1$ comme étant $\lambda_{t+1} = \sigma_{+,t}^2/C^2$. Puis, l'inégalité*

$$|f_\star(x) - f_{\lambda_{t+1},t}(x)| \leqslant \sqrt{\frac{k_{\lambda_{t+1},t}(x,x)}{\lambda_{t+1}}} B_{\lambda_{t+1},t}(\delta) \qquad \text{où}$$
$$B_{\lambda_{t+1},t}(\delta) = \sqrt{\lambda_{t+1}} C + \sigma_{+,t}\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda_-)} \tag{3.7}$$

*est valide avec probabilité supérieure à $1-4\delta$ simultanément pour tout $x \in \mathcal{X}$ et $t \geqslant 0$.*

*Démonstration.* Soit l'événement $E_f$ correspondant à la situation

$$|f_\star(x) - f_{\boldsymbol{\lambda},t}(x)| \leqslant \sqrt{\frac{k_{\lambda_{t+1},t}(x,x)}{\lambda_{t+1}}} \Big[\sqrt{\lambda_{t+1}}\|f_\star\|_{\mathcal{K}} + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda_\star)}\Big]$$

simultanément pour tout $x \in \mathcal{X}$ et $t \geqslant 0$, et soit l'événement $E_{\boldsymbol{\lambda}}$ signifiant que $\lambda_t \geqslant \lambda_\star$ est vrai pour tout $t$. Nous décomposons

$$\mathbb{P}[E_f^c] = \mathbb{P}[E_f^c \cap E_{\boldsymbol{\lambda}}] + \mathbb{P}[E_f^c \cap E_{\boldsymbol{\lambda}}^c] \leqslant \mathbb{P}[E_f^c \cap E_{\boldsymbol{\lambda}}] + \mathbb{P}[E_{\boldsymbol{\lambda}}^c]$$

Par le théorème 1, nous avons $\mathbb{P}[E_f^c \cap E_{\boldsymbol{\lambda}}] \leqslant \delta$. Nous devons montrer que $\lambda_t \geqslant \lambda_\star$ pour tout $t \geqslant 1$ en ajustant $\lambda_t$ en suivant la procédure proposée. Regardons ce qui se produit à chaque temps $t$. En utilisant la procédure proposée, nous avons $\lambda_1 = \sigma_+^2/C^2 \geqslant \lambda_\star$. Nous avons ensuite

$$\begin{aligned}
\sigma_{-,1}(\lambda_1) &\leqslant \sigma \quad \text{L1} \\
\sigma_{+,1}(\lambda_1, \sigma_{-,1}(\lambda_1)^2/C^2) &\geqslant \sigma \quad \text{U1} \\
\to \lambda_2 = \sigma_{+,1}(\lambda_1, \sigma_{-,1}(\lambda_1)^2/C^2)^2/C^2 &\geqslant \lambda_\star \\
\sigma_{-,2}(\lambda_2) &\leqslant \sigma \quad \text{L2} \\
\sigma_{+,2}(\lambda_2, \sigma_{-,2}(\lambda_2)^2/C^2) &\geqslant \sigma \quad \text{U2} \\
\to \lambda_3 = \sigma_{+,2}(\lambda_2, \sigma_{-,2}(\lambda_2)^2/C^2)^2/C^2 &\geqslant \lambda_\star \\
&\dots
\end{aligned}$$

tel que $E_{\boldsymbol{\lambda}}$ se réalise si les étapes L1, U1, L2, U2, ... sont vraies simultanément. Ainsi, $\mathbb{P}[E_{\boldsymbol{\lambda}}^c]$ est borné par la probabilité que ces étapes ne se soient pas vraies simultanément. Suivant le théorème 4, nous avons $\mathbb{P}[E_{\boldsymbol{\lambda}}^c] \leqslant 3\delta$, donc $\mathbb{P}[E_f^c] \leqslant 4\delta$. Naturellement, sous l'événement $E_{\boldsymbol{\lambda}}$, nous avons $\sigma_{+,t} \geqslant \sigma$ et $\lambda_- \leqslant \lambda_\star$. Ainsi, étant donné $C \geqslant \|f_\star\|_{\mathcal{K}}$, nous avons

$$\sqrt{\lambda_{t+1}}\|f_\star\|_{\mathcal{K}} + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda_\star)} \leqslant \sqrt{\lambda_{t+1}} C + \sigma_{+,t}\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda_-)}.$$

□

**Remarque 12.** *On observe que la rigueur des estimateurs du bruit rapportés au lemme 4 dépendent du paramètre de régularisation $\lambda$ utilisé pour calculer $\tilde{\sigma}_{-,t}$ et $\tilde{\sigma}_{+,t}$. Puisque $\sigma^2/C^2 \leqslant \lambda_t \leqslant \sigma_+^2/C^2$ est vrai avec grande probabilité par construction, l'utilisation de $\lambda_t$ devrait mener à l'obtention d'un estimateur plus serré qu'avec une régularisation fixe $\lambda = \sigma_+^2/C^2$. Des résultats numériques présentés à la section 3.6.2 appuient cette intuition.*

**Lemme 5** (Borne déterministe sur l'intervalle de confiance)**.** *Sous l'hypothèse qu'une constante* $0 < \sigma_- < \sigma$ *est disponible telle que* $\sigma_{-,t} \geqslant \sigma_-$ *est vrai pour tout* $t$*, alors pour* $1 \leqslant t \leqslant T$*,*

$$B_{\lambda_t,t-1}(\delta) \leqslant \sigma_+\left(1 + \sqrt{2\ln(1/\delta) + 2\gamma_T(\sigma_-^2/C^2)}\right).$$

*Rappelons que* $\gamma_t(\lambda) = \frac{1}{2}\sum_{s=1}^{t} \ln\left(1 + \frac{1}{\lambda}k_{\lambda,s-1}(x_s, x_s)\right)$ *dénote le gain d'information (voir la section 1.2.5).*

*Démonstration.* Il s'agit simplement de remplacer $\sqrt{\lambda_t} \leqslant \sigma^+/C$, $\sigma_{+,t} \leqslant \sigma_+$ et $\lambda_- \geqslant \sigma_-^2/C^2$. □

**Remarque 13.** *Les termes* $\sigma_+$ *et* $\gamma_t(\sigma_{t,-}^2/C^2)$ *peuvent respectivement être remplacés par les quantités plus raffinées* $\sigma + \mathcal{O}(1/\sqrt{t})$ *et* $\gamma_t(\sigma^2/C^2) + \mathcal{O}(1/\sqrt{t})$ *grâce à l'ordre des intervalles de confiance sur les estimateurs de la variance (voir la remarque 11).*

## 3.4 Kernel TS

Similairement à KernelUCB (Valko et al., 2013), qui étend UCB au cas RKHS, nous introduisons Kernel TS, qui étend TS au cas RKHS, utilisant la régression à noyau pour maintenir une distribution a posteriori sur $f_\star$. À chaque épisode, l'algorithme calcule la moyenne et la covariance postérieures étant données les observations obtenues précédemment. Kernel TS échantillonne ensuite une fonction de la distribution résultante et sélectionne le point $x_t$ à observer comme étant celui maximisant celle-ci. En pratique, l'évaluation d'une fonction échantillonnée de la distribution a posteriori (fournie par la régression à noyau) peut être approximée par l'échantillonnage ponctuel d'une distribution normale multivariée paramétrée par la moyenne et la covariance postérieures. On introduit donc la discrétisation $\mathbb{X} \subset \mathcal{X}$, utilisée pour l'évaluation ponctuelle de $\tilde{f}_t$, sur laquelle l'optimisation est effectuée. L'algorithme 7 présente la procédure résultante de Kernel TS basée sur la régression à noyau adaptative et l'estimation empirique du bruit (théorème 1 et corollaire 1).

**Remarque 14.** *En pratique, la discrétisation est supposée suffisamment fine pour que son impact sur le résultat de l'optimisation soit considéré comme marginal.*

**Remarque 15.** *L'algorithme ne connaît pas la vraie variance du bruit* $\sigma^2$ *; il utilise plutôt une borne supérieure* $\sigma_{+,t-1}^2$*.*

Grâce à quelques adaptations attentives, mais simples, de la démarche d'Agrawal and Goyal (2014), nous obtenons la borne suivante sur le regret.

**Théorème 2** (Kernel TS avec estimation empirique de la variance du bruit)**.** *Sous l'hypothèse que l'écart de sous-optimalité* $R = \max_{x\in\mathbb{X}}\left(f_\star(\star) - f_\star(x)\right)$ *est fini, alors le regret de Kernel*

**Algorithme 7** Kernel TS

Paramètres : discrétisation $\mathbb{X} \subset \mathcal{X}$, bornes $\sigma_-^2 \leqslant \sigma^2$ et $\sigma_+^2 \geqslant \sigma^2$ sur la variance du bruit, borne $C \geqslant \|f_\star\|_{\mathcal{K}}$ sur la norme de la fonction et noyau $k$ associé à $\mathcal{K}$, procédure de calcul d'une inflation de la variance $v_t^2$ pour tout $t$.

initialiser $\sigma_{-,0} = \sigma_-$ et $\sigma_{+,0} = \sigma_+$
**for all** épisode $t \geqslant 1$ **do**
  calculer $\lambda_t = \sigma_{+,t-1}^2 / C^2$
  calculer la moyenne postérieure $\widehat{\mathbf{f}}_{t-1} = \big(f_{\lambda_t,t-1}(x)\big)_{x\in\mathbb{X}}$
  calculer la covariance postérieure $\widehat{\mathbf{\Sigma}}_{t-1} = \frac{\sigma_{+,t-1}^2}{\lambda_t}\big(k_{\lambda_t,t-1}(x,x')_{x,x'\in\mathbb{X}}\big)$
  calculer $v_t^2$
  échantillonner $\tilde{f}_t \sim \mathcal{N}(\widehat{\mathbf{f}}_{t-1}, v_t^2\widehat{\mathbf{\Sigma}}_{t-1})$
  jouer $x_t = \arg\max_{x\in\mathbb{X}} \tilde{f}_t(x)$
  observer $y_t = f_\star(x_t) + \xi_t$
  calculer $\sigma_{-,t} = \max\{\sigma_{-,t}(\lambda_t), \sigma_{-,t-1}\}$ et $\lambda_- = \sigma_{-,t}^2/C^2$
  calculer $\sigma_{+,t} = \min\{\sigma_{+,t}(\lambda_t, \lambda_-), \sigma_{+,t-1}\}$
**end for**

*TS (algorithme 7) avec* $v_t = \frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma_{+,t-1}}$ *et* $B_{\lambda_t,t-1}$ *donné par le corollaire 1 après* $T$ *épisodes est borné par*

$$\mathfrak{R}(T) \leqslant C_{1,T}\Bigg(\sum_{t=1}^{T}\sqrt{\frac{k_{\lambda_t,t-1}(x_t,x_t)}{\lambda_t}}B_{\lambda_t,t-1}(\delta/4)\Bigg) + C_2 R\sqrt{T\ln(1/\delta)} + 4\pi e R\delta,$$

*avec probabilité* $1-3\delta$ *pour tout* $T \geqslant 0$, *où* $C_{1,T} = (4\sqrt{\pi e}+1)\Bigg(1+\sqrt{2\ln\Big(\frac{T(T+1)|\mathbb{X}|}{\sqrt{\pi}\delta}\Big)}\Bigg)$ *et* $C_2 = \sqrt{8\pi e}(1+\delta\sqrt{4\pi e})$. *Plus spécifiquement, nous avons*

$$\begin{aligned}\mathfrak{R}(T) &\leqslant C_{1,T}\frac{\sigma_+}{\sigma}\Big(1+\sqrt{2\ln(4/\delta)+2\gamma_T(\sigma_-^2/C^2)}\Big)C\sqrt{T\frac{2\gamma_T(\sigma^2/C^2)}{\ln(1+C^2/\sigma^2)}}\\ &\quad + C_2 R\sqrt{T\ln(1/\delta)} + 4\pi e R\delta\end{aligned}$$

*avec la même probabilité. La borne sur le regret de Kernel TS est donc d'ordre* $\mathcal{O}(C\sqrt{T\ln(T|\mathbb{X}|)}\gamma_T(\sigma^2/C^2))$ *avec probabilité* $1-3\delta$. *Rappelons que* $\gamma_t(\lambda) = \frac{1}{2}\sum_{s=1}^{t}\ln\big(1+\frac{1}{\lambda}k_{\lambda,s-1}(x_s,x_s)\big)$ *dénote le gain d'information (voir la section 1.2.5). La procédure d'analyse est détaillée à la section suivante.*

**Gonflement de la variance** Bien que Kernel TS tel que décrit par l'algorithme 7 soit supporté par la théorie, il peut sembler contre-intuitif en comparaison avec l'algorithme de TS *traditionnel*. En effet, l'analyse de Kernel TS suggère que la distribution a posteriori n'est pas suffisante pour garantir l'*optimisme* nécessaire, alors que ce n'est pas le cas traditionnellement avec TS (par exemple, voir les algorithmes 1 et 2). Une extension alternative du TS traditionnel aux bandits structurés en RKHS pourrait consister à échantillonner $\tilde{f}_t$ directement de $\mathcal{N}(\widehat{\mathbf{f}}_{t-1}, \widehat{\mathbf{\Sigma}}_{t-1})$, c'est-à-dire de la distribution postérieure sur $f_\star$. Cette version alternative

sans-inflation correspond à utiliser l'algorithme 7 avec $v_t = 1$. Des résultats expérimentaux présentés à la section 3.6.4 appuient cette intuition.

**Convergence de Kernel TS** Concrètement, Kernel TS implique les paramètres suivants : une borne supérieure $C$ sur la norme de la fonction, $\|f_\star\|_{\mathcal{K}}$ ; une borne supérieure $\sigma_+$ et inférieure $\sigma_-$ sur la variance du bruit, $\sigma$. Ces derniers affectent la convergence de l'algorithme à travers différents aspects. Plus spécifiquement, les analyses du regret de TS montrent que la convergence de TS dépend directement de la vitesse de convergence de l'estimateur de $f_\star$ (voir la section 1.1.3). De plus, par définition du gain d'information, ce dernier est naturellement lié à la nécessité d'explorer (voir la section 1.2.5). En effet, le compromis entre l'exploration et l'exploitation s'articule entre le gain potentiel en termes de récompense et le gain potentiel en termes d'information.

Tel que discuté précédemment (voir la section 3.3), le fait de considérer la norme de la fonction cible $f_\star$ dans la régularisation permet au modèle de régression de s'adapter à la complexité de $f_\star$. En utilisant une borne supérieure $C$ pour ajuster $\lambda_t = \sigma^2_{+,t-1}/C^2$, la lâcheté de $C$ implique naturellement une régularisation *trop* faible, laquelle mène à des modèles prédictifs surajustés (*overfitting*). Le partage d'information entre les observations se trouve alors limité, ce qui se traduit par une réduction de la vitesse avec laquelle les intervalles de confiance se resserrent autour de $f_{\lambda_t,t-1}$, ainsi qu'une augmentation du gain d'information $\gamma_t(\lambda_t)$ (voir la section 1.2.5).

Similairement, la borne supérieure $\sigma_+$ affecte l'estimateur $\sigma_{+,t}$, lequel a un impact sur la régularisation. Une borne plus lâche implique plus de chemin à parcourir avant que $\sigma_{+,t}$ converge vers $\sigma$. Des expériences empiriques à la section 3.6.2 illustrent cet effet. La borne inférieure $\sigma_-$ affecte quand à elle l'estimateur $\sigma_{-,t}$. Celui-ci a un impact direct sur la borne inférieure $\lambda_-$ sur la régularisation, laquelle est utilisée pour le calcul de $\sigma_{+,t}$. La lâcheté de $\sigma_-$ mène donc à une surestimation de $\sigma_{+,t}$. La régularisation étant directement proportionnelle à $\sigma_{+,t}$, une surestimation de la variance du bruit implique donc une régularisation trop importante, des modèles prédictifs trop généraux. L'information est alors partagée lorsqu'elle ne le devrait pas, ce qui se traduit également par une réduction de la vitesse avec laquelle les intervalles de confiance se resserrent autour de $f_{\lambda_t,t-1}$.

On bénéficie donc naturellement des bornes initiales les plus serrées possible pour favoriser la convergence de Kernel TS. Cependant, en dehors de la configuration initiale, la performance de l'algorithme demeure limitée par la difficulté du problème. Il est évident qu'un grand bruit rend l'apprentissage plus difficile, de même qu'une fonction complexe dans laquelle le partage d'information à travers les observations est limité. La dimensionnalité du problème, combiné à la complexité de la fonction, joue également un rôle non négligeable dans la convergence des approches basées sur la régression à noyau. Son effet est représenté par le gain d'information (voir la section 1.2.5).

## 3.5 Analyse théorique

Nous présentons ici l'analyse détaillée permettant d'obtenir la borne sur le regret de Kernel TS (théorème 2), soit l'algorithme 7 avec $v_t = \frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma_{+,t-1}}$ et $B_{\lambda_t,t-1}$ donné par le Corollaire 1. Nous suivons de près la technique de preuve d'Agrawal and Goyal (2014) tout en tentant de la clarifier et de la simplifier. L'idée générale consiste à diviser les actions en deux groupes : les actions *saturées* et les actions *non saturées*. Le premier groupe désigne les actions pour lesquelles les échantillons $\tilde{f}_t$ ont une faible probabilité de dominer $f_\star(\star)$ alors que le second désigne l'autre cas. Ce concept est lié à l'*optimisme* (Abeille and Lazaric, 2016), c'est-à-dire la probabilité d'échantillonner une valeur qui est plus élevée que l'optimum.

Soient les événements $\widehat{E}_t$ et $\tilde{E}_t$ représentant respectivement les situations où $\widehat{\mathbf{f}}_{t-1}$ et $\tilde{f}_t$ sont concentrées autour de leurs espérances respectives. Plus précisément, pour un niveau de confiance $\delta$ donné, nous introduisons

$$\widehat{E}_{t,\delta} = \{|f_\star(x) - f_{\lambda_t,t-1}(x)| \leqslant \widehat{C}_{t,\delta}(x) \quad \forall x \in \mathcal{X}\}$$
$$\tilde{E}_{t,\delta} = \{|f_{\lambda_t,t-1}(x) - \tilde{f}_t(x)| \leqslant \tilde{C}_{t,\delta}(x) \quad \forall x \in \mathcal{X}\},$$

pour des quantités $\widehat{C}_{t,\delta}(x)$ et $\tilde{C}_{t,\delta}(x)$ à définir. La démarche de preuve repose sur le contrôle de l'occurrence de ces deux événements, puis sur le contrôle du regret instantané lorsque ces événements se produisent.

### 3.5.1 Contrôle des événements

#### Contrôle de $\widehat{E}_{t,\delta}$

Par le théorème 1, nous avons que l'écart entre la moyenne postérieure et la fonction cible est borné par

$$\widehat{C}_{t,\delta}(x) = \sqrt{\frac{k_{\lambda_t,t-1}(x,x)}{\lambda_t}} B_{\lambda_t,t-1}(\delta/4),$$

avec probabilité supérieure à $1-\delta$. Ainsi, l'événement $\widehat{E}_{t,\delta}$ est contrôlé de la manière suivante :

$$\mathbb{P}[\widehat{E}_{t,\delta} \ \forall t \geqslant 1] \geqslant 1 - \delta.$$

#### Contrôle de $\tilde{E}_{t,\delta}$

Rappelons que les évaluations ponctuelles de $\tilde{f}_t$ correspondent à un ensemble de variables échantillonnées d'une distribution normale multivariée. Ainsi, chaque point évalué peut être considéré de manière indépendante comme une variable échantillonnée d'une distribution normale, $\tilde{f}_t(x)|\mathcal{H}_{t-1} \sim \mathcal{N}(f_{\lambda_t,t-1}(x), v_t^2 \frac{\sigma^2_{+,t}}{\lambda_t} k_{\lambda_t,t-1}(x,x))$. Alors, par l'union bound sur tous $x \in \mathbb{X}$ de la concentration des variables normales (lemme 8, annexe A), nous avons

$$\mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}] \leqslant \sum_{x \in \mathbb{X}} \frac{1}{\sqrt{\pi} z_x} e^{-z_x^2/2}$$

à condition que $z_x = \frac{\tilde{C}_{t,\delta}(x)}{v_t\sqrt{\frac{\sigma^2_{+,t}}{\lambda_t}k_{\lambda_t,t-1}(x,x)}} \geqslant 1$ pour tout $x \in \mathbb{X}$. Cela motive la définition suivante,

$$\tilde{C}_{t,\delta}(x) = c_{t,\delta}v_t\sqrt{\frac{\sigma^2_{+,t}}{\lambda_t}k_{\lambda_t,t-1}(x,x)}$$

pour une séquence $(c_{t,\delta})_{t\geqslant 1}$ bien choisie. Plus spécifiquement, pour

$$c_{t,\delta} = \max\left\{\sqrt{2\ln\left(\frac{t(t+1)|\mathbb{X}|}{\sqrt{\pi}\delta}\right)}, 1\right\},$$

nous avons

$$\mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}\ \forall t \geqslant 1] \leqslant \sum_{t\geqslant 1}\frac{|\mathbb{X}|}{\sqrt{\pi}c_{t,\delta}}e^{-c^2_{t,\delta}/2} = \sum_{t\geqslant 1}\frac{\delta}{c_{t,\delta}t(t+1)} \leqslant \sum_{t\geqslant 1}\frac{\delta}{t(t+1)} = \delta$$

tel que

$$\mathbb{P}[\tilde{E}_{t,\delta}\ \forall t \geqslant 1] \geqslant 1-\delta.$$

### 3.5.2 Contrôle du regret instantané

Notons dans un premier temps que, sous l'occurrence simultanée des événements $\widehat{E}_{t,\delta}$ et $\tilde{E}_{t,\delta}$, l'écart entre la fonction échantillonnée et la fonction cible pour tout $x \in \mathcal{X}$ est contrôlé par

$$\begin{aligned}
\left|f_\star(x) - \tilde{f}_t(x)\right| &\leqslant |f_\star(x) - f_{\lambda_t,t-1}(x)| + \left|f_{\lambda_t,t-1}(x) - \tilde{f}_t(x)\right| \\
&\leqslant \widehat{C}_{t,\delta}(x) + \tilde{C}_{t,\delta}(x) \\
&= \sqrt{\frac{k_{\lambda_t,t-1}(x,x)}{\lambda_t}}(B_{\lambda_t,t-1}(\delta/4) + c_{t,\delta}v_t\sigma_{+,t-1}) \\
&= s_{\lambda_t,t-1}(x)\Big(\underbrace{\frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma} + c_{t,\delta}v_t\frac{\sigma_{+,t-1}}{\sigma}}_{g_t(\delta)}\Big).
\end{aligned} \tag{3.8}$$

L'avant-dernière égalité utilise les définitions de $\widehat{C}_{t,\delta}(x)$ et $\tilde{C}_{t,\delta}(x)$ introduites précédemment. La dernière égalité utilise la définition de la variance postérieure sur la moyenne donnée par le modèle de régression à noyau avec régularisation adaptative (théorème 1).

Nous divisons alors les actions en deux groupes : celles pour lesquelles la valeur échantillonnée a une faible probabilité de dominer $f_\star(\star)$ et les autres.

#### Les actions saturées

Une action $x$ est dite saturée si $\tilde{f}_t(x)$ a une faible probabilité de dominer $f_\star(\star)$, ce qui se produit lorsque l'écart de sous-optimalité est supérieur à l'erreur d'échantillonnage. Nous définissons donc l'ensemble des actions saturées comme

$$\mathcal{S}_{t,\delta} = \Big\{x \in \mathbb{X} : f_\star(\star) - f_\star(x) > s_{\lambda_t,t-1}(x)g_t(\delta)\Big\}.$$

On remarque que, par définition, $\star \notin \mathcal{S}_{t,\delta}$ pour tout $t$. Considérons également l'action non saturée minimisant l'erreur d'échantillonnage,

$$x_{\mathcal{S},t} \overset{\text{def}}{=} \operatorname*{arg\,min}_{x \notin \mathcal{S}_{t,\delta}} s_{\lambda_t,t-1}(x),$$

et rappelons que la stratégie de Kernel TS consiste à sélectionner $x_t = \arg\max_{x \in \mathbb{X}} \tilde{f}_t(x)$. Ainsi, sous l'occurrence de l'événement $\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}$,

$$\begin{aligned} f_\star(\star) - f_\star(x_t) &= f_\star(\star) - f_\star(x_{\mathcal{S},t}) + f_\star(x_{\mathcal{S},t}) - f_\star(x_t) \\ &\leqslant s_{\lambda_t,t-1}(x_{\mathcal{S},t}) g_t(\delta) + f_\star(x_{\mathcal{S},t}) - \tilde{f}_t(x_{\mathcal{S},t}) + \underbrace{\tilde{f}_t(x_{\mathcal{S},t}) - \tilde{f}_t(x_t)}_{\leqslant 0} + \tilde{f}_t(x_t) - f_\star(x_t) \\ &\leqslant 2 s_{\lambda_t,t-1}(x_{\mathcal{S},t}) g_t(\delta) + s_{\lambda_t,t-1}(x_t) g_t(\delta). \end{aligned} \tag{3.9}$$

La première inégalité utilise le fait que l'action $x_{\mathcal{S},t}$ est non saturée. La dernière inégalité utilise la borne sur l'écart entre la fonction échantillonnée et la fonction cible (équation 3.8) pour les actions $x_{\mathcal{S},t}$ et $x_t$.

Nous relions $s_{\lambda_t,t-1}(x_{\mathcal{S},t})$ à $s_{\lambda_t,t-1}(x_t)$ en utilisant la définition de $x_{\mathcal{S},t}$, par laquelle nous avons

$$\begin{aligned} \mathbb{E}\big[s_{\lambda_t,t-1}(x_t) | \mathcal{H}_{t-1}\big] &\geqslant \mathbb{E}\big[s_{\lambda_t,t-1}(x_t) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\} | \mathcal{H}_{t-1}\big] \\ &\geqslant \mathbb{E}\big[s_{\lambda_t,t-1}(x_{\mathcal{S},t}) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\} | \mathcal{H}_{t-1}\big] \\ &= s_{\lambda_t,t-1}(x_{\mathcal{S},t}) \mathbb{P}[x_t \notin \mathcal{S}_{t,\delta} | \mathcal{H}_{t-1}], \end{aligned}$$

donc

$$s_{\lambda_t,t-1}(x_{\mathcal{S},t}) \leqslant \frac{\mathbb{E}\big[s_{\lambda_t,t-1}(x_t) | \mathcal{H}_{t-1}\big]}{\mathbb{P}[x_t \notin \mathcal{S}_{t,\delta} | \mathcal{H}_{t-1}]}.$$

Rappelons que $f_\star(\star) - f_\star(x_t) \leqslant R$, où $R = \max_{x \in \mathcal{X}} \big[f_\star(\star) - f_\star(x)\big] < \infty$. De plus, par définition des actions saturées, $(f_\star(\star) - f_\star(x_t)) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\} \leqslant s_{\lambda_t,t-1}(x_t) g_t(\delta) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\}$. Nous avons donc

$$\begin{aligned} f_\star(\star) - f_\star(x_t) &= (f_\star(\star) - f_\star(x_t)) \mathbb{I}\{x_t \in \mathcal{S}_{t,\delta}\} + (f_\star(\star) - f_\star(x_t)) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\} \\ &\leqslant \min\Big\{2 s_{\lambda_t,t-1}(x_{\mathcal{S},t}) g_t(\delta) + s_{\lambda_t,t-1}(x_t) g_t(\delta), R\Big\} \mathbb{I}\{x_t \in \mathcal{S}_{t,\delta}\} \\ &\qquad + s_{\lambda_t,t-1}(x_t) g_t(\delta) \mathbb{I}\{x_t \notin \mathcal{S}_{t,\delta}\} \\ &\leqslant \min\Big\{2 s_{\lambda_t,t-1}(x_{\mathcal{S},t}) g_t(\delta), R\Big\} \mathbb{I}\{x_t \in \mathcal{S}_{t,\delta}\} + s_{\lambda_t,t-1}(x_t) g_t(\delta) \\ &\leqslant \frac{\mathbb{E}\Big[\min\Big\{2 s_{\lambda_t,t-1}(x_t) g_t(\delta), R\Big\} | \mathcal{H}_{t-1}\Big]}{\mathbb{P}[x_t \notin \mathcal{S}_{t,\delta} | \mathcal{H}_{t-1}]} \mathbb{I}\{x_t \in \mathcal{S}_{t,\delta}\} + s_{\lambda_t,t-1}(x_t) g_t(\delta), \end{aligned} \tag{3.10}$$

où la première inégalité utilise l'équation 3.9. On remarque que le regret instantané dépend de la probabilité de sélectionner une action non saturée. Tel que discuté précédemment, une action peut-être non saturée parce que son écart de sous-optimalité est petit ou parce que

son erreur d'échantillonnage est grande. Le premier cas est évidemment intéressant parce qu'il n'implique qu'un petit regret instantané (exploitation). Le second cas est également intéressant parce qu'il implique une réduction de l'incertitude dans une région de l'espace méconnue (exploration).

**Contrôle de l'échantillonnage d'actions non saturées**

À ce point, nous savons que sous l'événement $\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}$, nous avons

$$\tilde{f}_t(x) \leqslant f_\star(x) + s_{\lambda_t,t-1}(x) g_t(\delta) \leqslant f_\star(\star) \quad \forall x \in \mathcal{S}_{t,\delta}.$$

Puisque $\star \notin \mathcal{S}_{t,\delta}$ par définition, nous savons que si l'échantillon en $\star$ domine les échantillons aux points saturés, alors une action non saturée sera assurément sélectionnée :

$$\{\tilde{f}_t(\star) > \tilde{f}_t(x) \quad \forall x \in \mathcal{S}_{t,\delta}\} \subset \{x_t \notin \mathcal{S}_{t,\delta}\}.$$

La combinaison de ces deux propriétés nous permet de déduire que

$$\begin{aligned}
&\{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta} \\
&\quad \subset \{\exists x \in \mathcal{S}_{t,\delta} : \tilde{f}_t(\star) \leqslant \tilde{f}_t(x)\} \cap \{\forall x \in \mathcal{S}_{t,\delta}, \tilde{f}_t(x) \leqslant f_\star(\star)\} \\
&\quad \subset \{\tilde{f}_\star(\star) \leqslant f_\star(\star)\}.
\end{aligned}$$

De plus, en utilisant le fait que $\tilde{f}_t(x)|\mathcal{H}_{t-1} \sim \mathcal{N}(f_{\lambda_t,t-1}(x), v_t^2 \frac{\sigma_{+,t}^2}{\lambda_t} k_{\lambda_t,t-1}(x,x))$, nous avons

$$\begin{aligned}
&\{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta} \\
&\quad \subset \{\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star) \leqslant f_\star(\star) - f_{\lambda_t,t-1}(\star)\} \cap \widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta} \\
&\quad \subset \{\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star) \leqslant \widehat{C}_{t,\delta}(\star)\} \\
&\quad \subset \{|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| \leqslant \widehat{C}_{t,\delta}(\star)\},
\end{aligned}$$

donc

$$\begin{aligned}
\{|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star)\} \subset &\{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}^c_{t,\delta} \cap \tilde{E}_{t,\delta} \\
&\cup \{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}_{t,\delta} \cap \tilde{E}^c_{t,\delta} \\
&\cup \{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}^c_{t,\delta} \cap \tilde{E}^c_{t,\delta} \\
&\cup \{x_t \notin \mathcal{S}_{t,\delta}\}
\end{aligned}$$

et

$$\begin{aligned}
\{|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star)\} \cap \widehat{E}_{t,\delta} &\subset \{x_t \in \mathcal{S}_{t,\delta}\} \cap \widehat{E}_{t,\delta} \cap \tilde{E}^c_{t,\delta} \cup \{x_t \notin \mathcal{S}_{t,\delta}\} \cap \widehat{E}_{t,\delta} \\
&\subset \{x_t \notin \mathcal{S}_{t,\delta}\} \cup \tilde{E}^c_{t,\delta}.
\end{aligned}$$

Cela prouve donc que

$$\begin{aligned}
\mathbb{P}[x_t \notin \mathcal{S}_{t,\delta}|\mathcal{H}_{t-1}] &\geqslant \mathbb{P}[|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star), \widehat{E}_{t,\delta}] - \mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}] \\
&= \mathbb{P}[|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star)] \mathbb{I}\{\widehat{E}_{t,\delta}\} - \mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}]. \qquad (3.11)
\end{aligned}$$

En utilisant l'anti-concentration des variables normales (lemme 8, annexe A), nous avons

$$\mathbb{P}[|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star)] \geqslant \frac{1}{2\sqrt{\pi}z} e^{-z^2/2},$$

où $z = \frac{\widehat{C}_{t,\delta}(\star)}{v_t \sigma_{+,t-1}\sqrt{\frac{k_{\lambda_t,t-1}(\star,\star)}{\lambda_t}}} = \frac{B_{\lambda_t,t-1}(\delta/4)}{v_t \sigma_{+,t-1}}$ à condition que $z \geqslant 1$. En prenant

$$v_t = \frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma_{+,t-1}\sqrt{2\alpha_t \ln(\beta_t)}}$$

avec des constantes $\alpha_t$ et $\beta_t$ telles que $z = \sqrt{2\alpha_t \ln(\beta_t)} \geqslant 1$, nous obtenons

$$\mathbb{P}[|\tilde{f}_\star(\star) - f_{\lambda_t,t-1}(\star)| > \widehat{C}_{t,\delta}(\star)] \geqslant p_t \stackrel{\text{def}}{=} \frac{\beta_t^{-\alpha_t}}{2\sqrt{\pi}\sqrt{2\alpha_t \ln(\beta_t)}}. \tag{3.12}$$

En combinant les équations 3.10, 3.11 et 3.12, nous obtenons

$$\begin{aligned}
&(f_\star(\star) - f_\star(x_t))\mathbb{I}\{\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}\} \\
&\quad\leqslant \frac{\mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]}{p_t\mathbb{I}\{\widehat{E}_{t,\delta}\} - \mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}]}\mathbb{I}\{x_t \in \mathcal{S}_{t,\delta}\}\mathbb{I}\{\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}\} \\
&\qquad + s_{\lambda_t,t-1}(x_t)g_t(\delta)\mathbb{I}\{\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}\} \\
&\quad\leqslant \frac{\mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]}{p_t\mathbb{I}\{\widehat{E}_{t,\delta}\} - \mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}]} + s_{\lambda_t,t-1}(x_t)g_t(\delta) \\
&\quad\leqslant \mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]\left(\frac{1}{p_t} + \frac{\mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}]}{p_t^2}\right) + s_{\lambda_t,t-1}(x_t)g_t(\delta),
\end{aligned}$$

où la dernière inégalité utilise le fait que $\frac{1}{p-q} = \frac{1}{p} + \frac{p}{p(q-p)} \leqslant \frac{1}{p} + \frac{q}{p^2}$ pour $p \geqslant q$. Sachant que $\mathbb{P}[\tilde{E}^c_{t,\delta}|\mathcal{H}_{t-1}] \leqslant \frac{\delta}{c_{t,\delta}t(t+1)}$ (voir la section 3.5.1) et la définition de $p_t$, nous poursuivons

$$\begin{aligned}
&(f_\star(\star) - f_\star(x_t))\mathbb{I}\{\widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}\} \\
&\quad\leqslant \mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]\left(\sqrt{8\pi\alpha_t \ln(\beta_t)}\beta_t^{\alpha_t} + \delta\frac{8\pi\alpha_t \ln(\beta_t)\beta_t^{2\alpha_t}}{c_{t,\delta}t(t+1)}\right) \\
&\qquad + s_{\lambda_t,t-1}(x_t)g_t(\delta).
\end{aligned}$$

**Somme et concentration**

Nous effectuons maintenant la sommation sur $t \geqslant 1$ et obtenons que sous l'événement $\bigcap_{t\geqslant 1} \widehat{E}_{t,\delta} \cap \tilde{E}_{t,\delta}$ qui se produit avec probabilité supérieure à $1 - 2\delta$,

$$\begin{aligned}
\mathfrak{R}(T) \leqslant \sum_{t=1}^{T}\Bigg[&\mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]\left(\sqrt{8\pi\alpha_t \ln(\beta_t)}\beta_t^{\alpha_t} + \delta\frac{8\pi\alpha_t \ln(\beta_t)\beta_t^{2\alpha_t}}{c_{t,\delta}t(t+1)}\right) \\
&+ s_{\lambda_t,t-1}(x_t)g_t(\delta)\Bigg],
\end{aligned}$$

où $c_{t,\delta} = \max\Big\{\sqrt{2\ln\big(t(t+1)|\mathbb{X}|/(\sqrt{\pi}\delta)\big)}, 1\Big\}$ et les constantes $\alpha_t$, $\beta_t$ sont telles que $2\alpha_t \ln(\beta_t) \geqslant 1$ et $\sqrt{8\pi\alpha_t \ln(\beta_t)\beta_t^{\alpha_t}} \geqslant 1$. De plus, rappelons (voir l'équation 3.8) que

$$g_t(\delta) = \frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma} + c_{t,\delta} v_t \frac{\sigma_{+,t-1}}{\sigma} = \frac{B_{\lambda_t,t-1}(\delta/4)}{\sigma}\Big(1 + \frac{c_{t,\delta}}{\sqrt{2\alpha_t \ln(\beta_t)}}\Big).$$

Plus particulièrement, pour $\alpha_t = (2\ln(\beta_t))^{-1}$ et $\beta_t > 1$ (ce qui satisfait $1 \geqslant 1$ et $\sqrt{4\pi e} \geqslant 1$), nous obtenons

$$\mathfrak{R}(T) \leqslant \sum_{t=1}^{T}\left[\mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]\eta_t + s_{\lambda_t,t-1}(x_t)g_t(\delta)\right],$$

avec

$$\eta_t = \sqrt{4\pi e}\Big(1 + \delta\frac{\sqrt{4\pi e}}{c_{t,\delta}t(t+1)}\Big) \leqslant \sqrt{4\pi e}\Big(1 + \delta\frac{\sqrt{4\pi e}}{t(t+1)}\Big).$$

Il s'agit maintenant de relier la somme des termes $\mathbb{E}[s_{\lambda_t,t-1}(x_t)|\mathcal{H}_{t-1}]$ à la somme des termes $s_{\lambda_t,t-1}(x_t)$, pour $t \geqslant 1$. À cet effet, nous introduisons la variable aléatoire

$$X_t = \mathbb{E}\Big[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}|\mathcal{H}_{t-1}\Big]\eta_t - \min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}\eta_t.$$

Par définition, $\mathbb{E}[X_t] = 0$ et $|X_t| \leqslant R\eta_t$. En appliquant l'inégalité d'Azuma-Hœffding pour les martingales (lemme 11, annexe A), nous obtenons que pour tout $\delta \in (0,1)$, avec probabilité supérieure à $1-\delta$,

$$\sum_{t=1}^{T} X_t \leqslant \sqrt{2\sum_{t=1}^{T} R^2\eta_t^2 \ln(1/\delta)}.$$

Ainsi, avec probabilité supérieure à $1-3\delta$,

$$\begin{aligned}
\mathfrak{R}(T) &\leqslant \sum_{t=1}^{T}\left[\min\Big\{2s_{\lambda_t,t-1}(x_t)g_t(\delta), R\Big\}\eta_t + s_{\lambda_t,t-1}(x_t)g_t(\delta)\right] + \sqrt{2\sum_{t=1}^{T} R^2\eta_t^2\ln(1/\delta)} \\
&\leqslant (4\sqrt{\pi e}+1)\sum_{t=1}^{T} s_{\lambda_t,t-1}(x_t)g_t(\delta) + R\delta 4\pi e + R\sqrt{8\pi e\sum_{t=1}^{T}\Big(1+\delta\frac{\sqrt{4\pi e}}{t(t+1)}\Big)^2\ln(1/\delta)} \\
&\leqslant (4\sqrt{\pi e}+1)\sum_{t=1}^{T} s_{\lambda_t,t-1}(x_t)g_t(\delta) + R\delta 4\pi e + R\sqrt{8\pi e(1+\delta\sqrt{4\pi e})^2 T\ln(1/\delta)} \\
&\leqslant (4\sqrt{\pi e}+1)\sum_{t=1}^{T}\frac{k_{\lambda_t,t-1(x_t,x_t)}}{\lambda_t} B_{\lambda_t,t-1}(\delta/4)(1+c_{t,\delta}) \\
&\quad + R\delta 4\pi e + R\sqrt{8\pi e(1+\delta\sqrt{4\pi e})^2 T\ln(1/\delta)},
\end{aligned}$$

où la deuxième inégalité remplace $\eta_t$. Cela conclut la preuve du théorème 2 puisque $c_{t,\delta} \leqslant c_{T,\delta}$. Un résultat plus spécifique peut être obtenu en utilisant le lemme 5 ainsi que le lemme suivant.

**Lemme 6** (Somme des variances postérieures)**.** *Sous l'hypothèse que le noyau est borné par 1, c'est-à-dire* $\sup_{x\in\mathcal{X}} k(x,x) \leqslant 1$*, et pour n'importe quelle séquence* $\boldsymbol{\lambda}$ *telle que* $\lambda \geqslant \sigma^2/C^2$ *pour tout* $\lambda \in \boldsymbol{\lambda}$*, alors*

$$\sum_{t=1}^{T} s^2_{\lambda_t,t-1}(x_t) = \sigma^2 \sum_{t=1}^{T} \frac{1}{\lambda_t} k_{\lambda_t,t-1}(x_t,x_t) \leqslant \frac{2C^2}{\ln(1+C^2/\sigma^2)} \gamma_T(\sigma^2/C^2).$$

*Par exemple, la seconde hypothèse est satisfaite avec forte probabilité lorsque la séquence* $\boldsymbol{\lambda}$ *est construite avec l'équation 3.4.*

*Démonstration.* La démarche se base de près sur la preuve du lemme 7.1 de Srinivas et al. (2012). En utilisant $\min\{r,\alpha\} \leqslant (\alpha/\ln(1+\alpha))\ln(1+r)$ et le fait que $\min_{\lambda\in\boldsymbol{\lambda}} \lambda \geqslant \sigma^2/C^2$, nous obtenons

$$\begin{aligned}
\sum_{t=1}^{T} s^2_{\boldsymbol{\lambda},t-1}(x_t) &= \sigma^2 \sum_{t=1}^{T} \frac{1}{\lambda_t} k_{\lambda_t,t-1}(x_t,x_t) \\
&\leqslant \sigma^2 \sum_{t=1}^{T} \frac{C^2}{\sigma^2} k_{\sigma^2/C^2,t-1}(x_t,x_t) \\
&= \sigma^2 \sum_{t=1}^{T} \min\Big\{ \frac{C^2}{\sigma^2} k_{\sigma^2/C^2,t-1}(x_t,x_t), \frac{C^2}{\sigma^2} \Big\} \\
&\leqslant \frac{2C^2}{\ln(1+C^2/\sigma^2)} \gamma_T(\sigma^2/C^2).
\end{aligned}$$

En particulier, avec l'inégalité de Cauchy-Schwarz, nous obtenons

$$\sum_{t=1}^{T} \sqrt{\frac{k_{\lambda_t,t-1}(x_t,x_t)}{\lambda_t}} \leqslant \sqrt{T \frac{2C^2/\sigma^2}{\ln(1+C^2/\sigma^2)} \gamma_T(\sigma^2/C^2)}.$$

□

## 3.6 Évaluation empirique

Nous illustrons maintenant les résultats théoriques introduits dans ce chapitre à l'aide d'expérimentations synthétiques. Dans un premier temps, nous illustrons l'intervalle de confiance sur la régression à noyau présenté au lemme 2, puis nous illustrons les estimateurs empiriques du bruit présentés au lemme 4. Nous illustrons ensuite l'intervalle de confiance sur la régression à noyau avec régularisation adaptative basée sur l'estimation empirique du bruit, soit le corollaire 1. Nous présentons finalement une évaluation de l'algorithme Kernel TS proposé pour les bandits structurés en RKHS.

Les expériences sont effectuées avec la fonction $f_\star(x) = \theta_\star^\top \varphi(x)$ montrée par la figure 3.1. Cette fonction possède une norme $\|f_\star\|_{\mathcal{K}} = \|\theta_\star\|_2 = 2.06$ dans le RKHS induit par un noyau gaussien $k(x,x') = \varphi(x)^\top \varphi(x')$ (équation 1.26) de bande passante $\rho = 0.3$. Cette fonction a

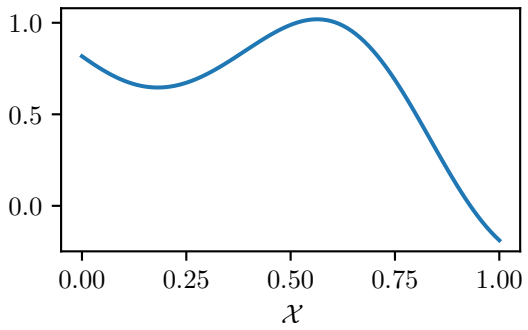



FIGURE 3.1 – Fonction test $f_\star$ utilisée dans les exemples synthétiques suivants.

été produite en générant directement un vecteur de paramètres $\theta_\star$ à l'aide d'une expansion de Taylor [4]. Nous considérons l'espace $\mathcal{X} = [0, 1]$ ainsi qu'un bruit d'écart-type $\sigma = 0.1$ sur les observations. Toutes les expériences qui suivent utilisent la borne supérieure $C = 5$ sur $\|f_\star\|_\mathcal{K}$ et la borne inférieure $\sigma_- = 0.01$ sur $\sigma$.

### 3.6.1 Intervalle de confiance sur la régression à noyau

Suite à l'introduction du lemme 2, il a été remarqué que la régularisation $\lambda = \sigma^2/C^2$ semblait appropriée étant donné l'intervalle de confiance sur la régression à noyau. Cela diffère de l'intuition bayésienne habituelle suggérant plutôt $\lambda = \sigma^2$. La figure 3.2 illustre les intervalles de confiance résultants du lemme 2 pour les deux paramètres de régularisation suggérés et avec $\delta = 0.1$, après différents nombres d'observations échantillonnées uniformément de $\mathcal{X}$. Rappelons que cet intervalle de confiance utilise la connaissance du vrai bruit ($\sigma = 0.1$). On observe que la régularisation suggérée par l'intervalle de confiance du lemme 2 permet d'obtenir une enveloppe de confiance plus serrée. Les expériences qui suivent utiliseront cette régularisation particulière.

### 3.6.2 Estimation empirique du bruit par régression à noyau

Nous illustrons maintenant la convergence des estimateurs $\sigma_{-,t} = \max\{\sigma_{-,t}(\lambda), \sigma_{-,t-1}\}$ et $\sigma_{+,t} = \min\{\sigma_{+,t}(\lambda, \lambda_-), \sigma_{+,t-1}\}$ calculés avec le lemme 4, où $\lambda_- = \sigma^2_{-,t}/C^2$ et $\delta = 0.1$. Les observations sont échantillonnées uniformément de $\mathcal{X}$. La remarque 12 suggère que la régularisation $\lambda = \sigma^2_{+,t-1}/C^2$ devrait permettre d'obtenir des estimateurs plus précis qu'une régularisation fixe $\lambda = \sigma^2_+/C^2$. La figure 3.3 montre que c'est en effet le cas, plus spécifiquement pour un nombre grandissant d'observations. On observe que l'utilisation d'une régularisation adaptative dans l'estimation empirique de la variance permet de converger à la même valeur, peu importe la borne supérieure initiale $\sigma_+$. Ce résultat est particulièrement intéressant lorsque $\sigma_+$ s'avère être une borne lâche sur $\sigma$.

4. Si $x \in \mathbb{R}^1$, la $i$-ième caractéristique d'un noyau gaussien $\varphi_i(x) = e^{-x^2/2\rho^2} \frac{x^{i-1}}{\rho^{i-1}\sqrt{(i-1)!}}$ pour $i \in \mathbb{N}_{>0}$ (Cotter et al., 2011).

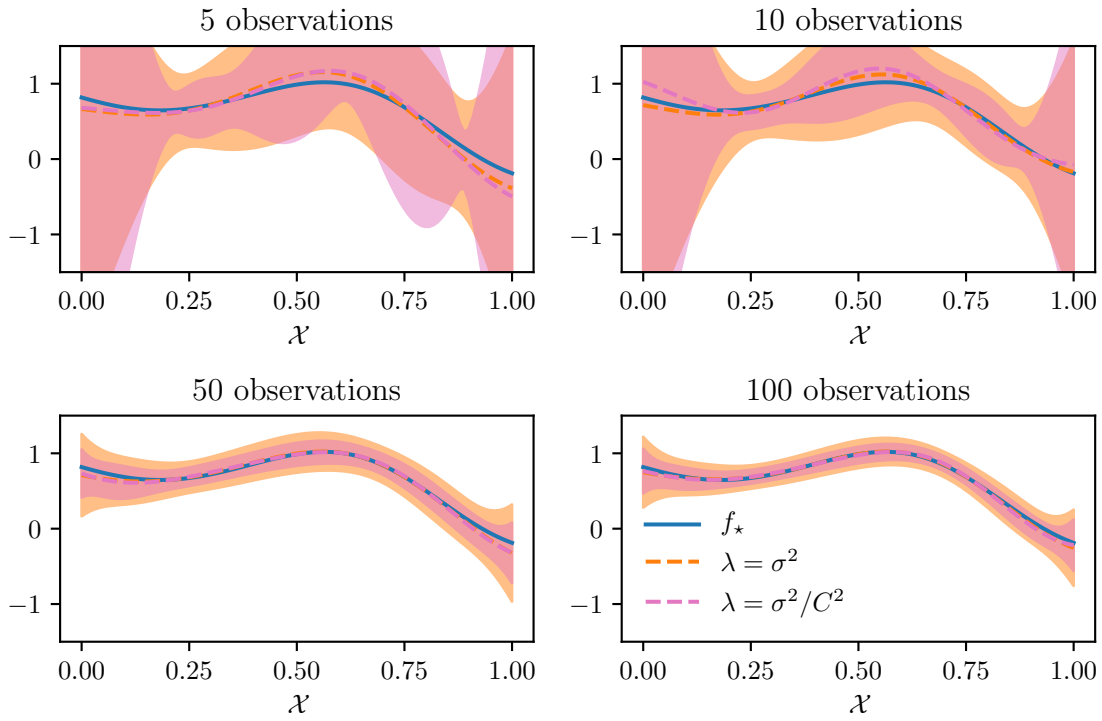
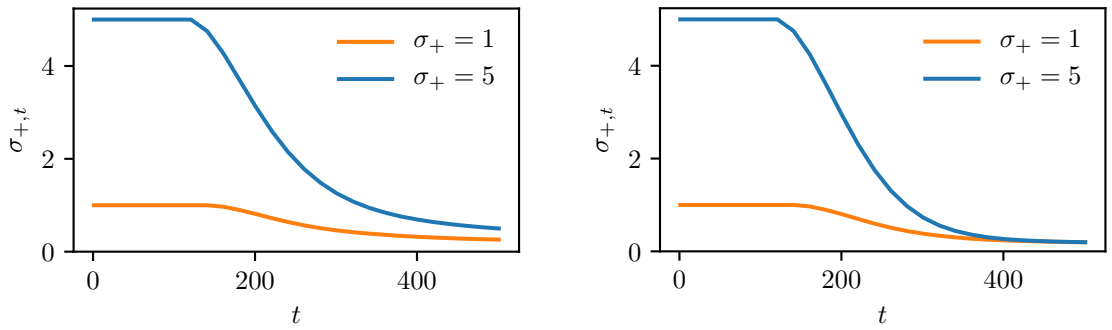



FIGURE 3.2 – Impact de la régularisation sur les intervalles de confiance (lemme 2).

En pratique, les estimateurs fournis par le lemme 4 pour $\sigma_+$ inconnu sont utiles même lorsqu'une borne $\sigma_+$ est disponible. Par exemple, la figure 3.4a montre l'évolution de la borne $\sigma_{+,t}$ avec et sans la connaissance d'une borne $\sigma_+$. On observe qu'après un certain nombre d'observations, la borne agnostique devient plus serrée. Nous suggérons donc, en pratique, de définir $\sigma_{+,t}(\lambda, \lambda_-)$ et $\sigma_{-,t}(\lambda)$ respectivement comme le minimum et le maximum entre la borne agnostique et celle utilisant la connaissance de $\sigma_+$. Autrement dit, nous considérons l'enveloppe de confiance la plus serrée autour de $\sigma$. La figure 3.4b montre les enveloppes de confiance autour de $\sigma$ obtenues avec cette procédure pour différentes valeurs de $\sigma_+$ (rappelons que $\sigma = 0.1$). On observe que la borne supérieure résultante pour $\sigma_+ = 1.0$ sur la figure 3.4b ne correspond pas au minimum entre les deux courbes sur la figure 3.4a. Cela s'explique par le fait que les estimateurs $\sigma_{+,t}$ et $\sigma_{-,t}$ dépendent de $\sigma_{+,t-1}$ et $\sigma_{-,t-1}$ à travers $\lambda$ et $\lambda_\star$. Les estimateurs (avec et sans $\sigma_+$) sont donc calculés en utilisant la même régularisation $\lambda_t = \sigma^2_{+,t-1}/C^2$ et $\lambda_- = \sigma^2_{-,t}/C^2$, soit la *meilleure* régularisation étant donné les estimateurs.

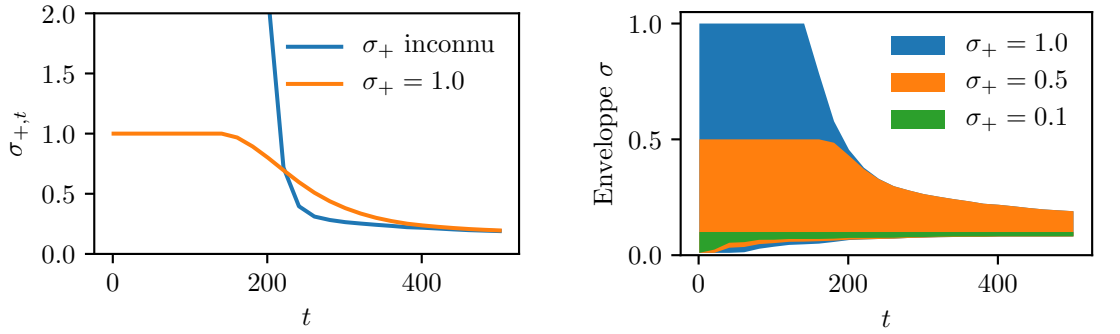
### 3.6.3 Régularisation adaptative

Nous utilisons maintenant les estimateurs empiriques du bruit pour ajuster automatiquement le paramètre de régularisation dans la régression à noyau, tel que décrit par le corollaire 1. Rappelons que nous considérons $\sigma_{-,0} = \sigma_-$, $\sigma_{+,0} = \sigma_+$ et $\lambda_1 = \sigma^2_+/C^2$. À chaque épisode $t \geqslant 1$, nous définissons $\lambda_{+,t} = \sigma^2_{+,t-1}/C^2$. Nous obtenons alors une nouvelle observation $y_t$ au point $x_t$. Nous calculons ensuite la borne inférieure sur le bruit, $\sigma_{-,t} = \max\{\sigma_{-,t}(\lambda_{t-1}), \sigma_{-,t-1}\}$,




(a) $\lambda = \sigma_+^2/C^2$ (b) $\lambda = \sigma_{+,t-1}^2/C^2$

FIGURE 3.3 – Impact de la régularisation sur l'estimation empirique du bruit (lemme 4).




(a) Convergence de l'estimateur (b) Enveloppe de confiance la plus serrée

FIGURE 3.4 – Impact de la connaissance de $\sigma_+$ sur l'estimation empirique du bruit (lemme 4) a) sans/avec connaissance de $\sigma_+$ et b) comme le minimum/maximum entre la borne agnostique et celle utilisant la connaissance de $\sigma_+$.

et définissons $\lambda_- = \sigma_{-,t}^2/C^2$. Nous calculons ensuite la borne supérieure sur le bruit, $\sigma_{+,t} = \min\{\sigma_{+,t}(\lambda_t, \lambda_-), \sigma_{+,t-1}\}$, lequel servira à calculer $\lambda_{+,t+1}$, et ainsi de suite. Nous obtenons alors l'intervalle de confiance donné par le corollaire 1. La figure 3.5 montre les intervalles de confiance résultants avec $\delta = 0.1$ et $\sigma_+ = 1$ (rappelons que $\sigma = 0.1$), après différents nombres d'observations échantillonnées uniformément de $\mathcal{X}$. Ces derniers sont comparés aux intervalles de confiance obtenus avec le lemme 2 pour une régularisation fixe $\lambda_t = \sigma_+^2/C^2$.

### 3.6.4 Kernel TS : Entre la théorie et la pratique

Nous présentons maintenant des résultats expérimentaux visant à illustrer la performance de Kernel TS pour des bandits aux actions structurées dans un RKHS. Nous optimisons toujours la fonction (inconnue) illustrée par la figure 3.1 avec un bruit $\sigma = 0.1$. Nous considérons l'espace $\mathbb{X}$ correspondant à une discrétisation linéaire de l'espace $\mathcal{X} = [0, 1]$ en 100 actions. Le but est de minimiser le pseudo-regret cumulatif (équation 3.1).

Nous évaluons Kernel TS donné par l'algorithme 7 avec garanties théoriques (théorème 2),

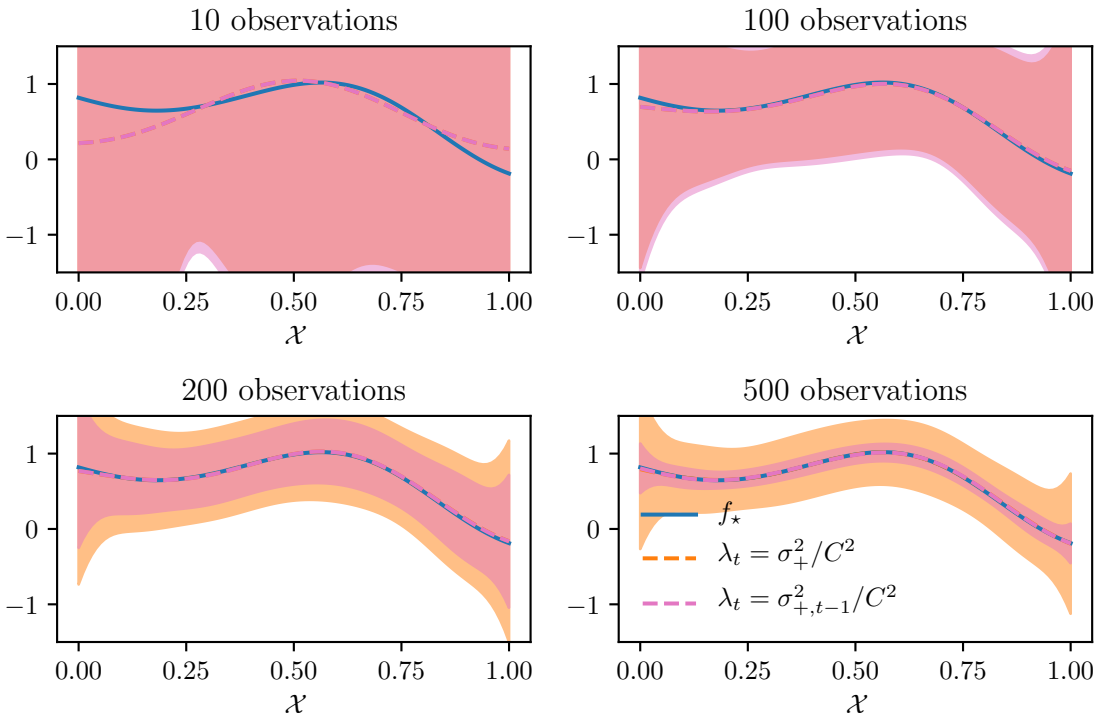



FIGURE 3.5 – Comparaison des intervalles de confiance avec régularisation fixe et adaptative.

c'est-à-dire avec $v_t = \frac{B_{\lambda_t,t-1}(\delta)}{\sigma_{+,t-1}}$, et Kernel TS sans inflation (donc sans garanties théoriques), c'est-à-dire avec $v_t = 1$. Kernel TS est comparé avec l'extension de l'approche classique UCB aux bandits structurés en RKHS, Kernel UCB (Valko et al., 2013). Plus précisément, nous considérons KernelUCB utilisant les intervalles de confiance du corollaire 1 pour supporter la régularisation adaptative. Spécifiquement, KernelUCB sélectionne l'action

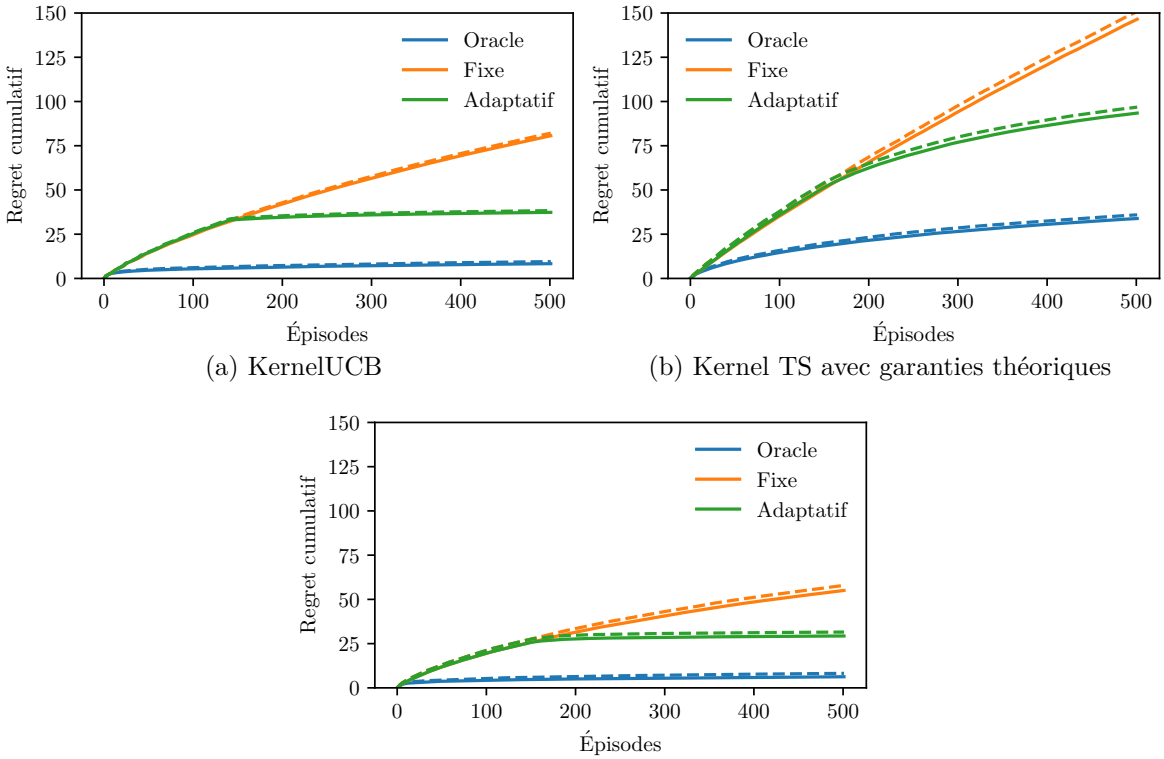
$$x_t = \arg\max_{x \in \mathbb{X}} \left[ f_{\lambda_t,t-1}(x) + \sqrt{\frac{k_{\lambda_t,t-1}(x,x)}{\lambda_t}} B_{\lambda_t,t-1}(\delta) \right], \tag{3.13}$$

où $B_{\lambda_t,t-1}$ est donné par l'équation 3.7.

**Remarque 16.** *L'équation 3.13 correspond à l'extension de KernelUCB classique (équation 3.3) pour une régularisation adaptative basée sur l'estimation empirique du bruit.*

Pour illustrer le bénéfice de la régularisation adaptative, Kernel TS (avec et sans inflation) et KernelUCB sont évalués dans trois configurations :

a) l'oracle, c'est-à-dire avec une régularisation fixe $\lambda_t = \sigma^2/C^2$, supposant $\sigma$ connu ;
b) la régularisation fixe $\lambda_t = \sigma_+^2/C^2$, c'est-à-dire le mieux que l'on puisse faire sans connaissance a priori de $\sigma$ ;
c) la régularisation adaptative ajustée avec le corollaire 1.




(a) KernelUCB

(b) Kernel TS avec garanties théoriques

(c) Kernel TS sans inflation

FIGURE 3.6 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart-type supérieur (ligne pointillée) des trois configurations de chaque algorithme.

Toutes les configurations utilisent la borne supérieure $C = 5$ sur $\|f_\star\|_{\mathcal{K}}$. KernelUCB utilise $\delta = 0.1/4$ et Kernel TS (théorique) utilise $\delta = 0.1/12$ de telle sorte que leur borne respective sur le regret tiennent avec une probaiblité supérieure à 0.9. Les expériences sont réalisées sur 500 épisodes et répétées 100 fois.

La figure 3.6 montre le pseudo-regret cumulatif, moyenné sur les 100 répétitions, pour chaque configuration de chaque algorithme. Ces résultats confirment que l'ajustement adaptatif de la régularisation basé sur une estimation empirique du bruit peut conduire à une amélioration majeure par rapport à l'utilisation d'une estimation fixe et peu précise du bruit : après une phase initiale d'amorçage, le regret de l'algorithme adaptatif est cumulé au même rythme que celui de l'oracle connaissant exactement le bruit. Le fait que KernelUCB performe mieux que Kernel TS avec garanties théoriques, ainsi que la différence marquée entre Kernel TS avec et sans inflation, suggère que l'inflation de la variance dans Kernel TS, comme encouragé par la théorie, n'est peut-être pas optimal en pratique. En effet, on remarque que Kernel TS sans inflation est beaucoup plus compétitif avec KernelUCB, similairement à ce qui est typiquement observé en bandits traditionnels. Une attention particulière devrait être accordée à cette question.

## 3.7 Discussion

Dans ce chapitre, nous avons abordé le problème de bandits structurés. Plus spécifiquement, nous avons considéré l'optimisation en-ligne de fonctions appartenant à un RKHS par le biais de la régression à noyau. Contrairement aux travaux précédents, nous avons considéré le problème dans lequel les observations de la fonction à optimiser sont bruitées, d'un bruit de variance inconnue. Nous avons débuté par l'extension du résultat de concentration de la régression à noyau avec régularisation fixe pour supporter une régularisation adaptative (théorème 1). À notre connaissance, aucun résultat de ce genre n'existe dans la littérature au moment de la rédaction de ce document. Puis, nous avons introduit une procédure utilisant des estimateurs empiriques de bruit pour ajuster automatiquement la régularisation en se basant sur les observations antérieures (corollaire 1). Cette procédure est complètement adaptative et possède des intervalles de confiance explicites. Nous avons ensuite proposé une adaptation de l'algorithme de TS au cas des bandits structurés en RKHS, Kernel TS, pour lequel nous fournissons des garanties de convergence théoriques (théorème 2).

Les outils de régression à noyau avec régularisation variable ainsi que la procédure d'adaptation de la régularisation introduits dans ce chapitre lèvent l'hypothèse sur la connaissance a priori de la variance du bruit. Ils permettent ainsi de maintenir les garanties de convergence d'algorithmes basés sur la régression à noyau dans des contextes d'application réalistes. Cela contribue directement à réduire l'écart entre la théorie et la pratique. Une approche basée sur Kernel TS est utilisée pour une application concrète au chapitre 5. Cependant, tout comme les travaux antérieurs (Srinivas et al., 2010; Valko et al., 2013), les garanties sur le regret de Kernel TS présentées dans cette thèse supposent que le noyau utilisé dans la régression est *adapté* à la fonction cible. Plus spécifiquement encore, que la fonction cible appartient au RKHS du noyau. La possibilité de lever cette hypothèse, ou du moins d'encadrer les répercussions de son non-respect, pourrait permettre d'étendre davantage les garanties théoriques aux applications réelles.

Tout comme la variante de TS pour les bandits avec structure linéaire (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016), l'analyse théorique de Kernel TS suggère une inflation de la covariance alors que les résultats expérimentaux suggèrent autrement. Kernel TS sans gonflement de la variance présente une performante similaire à KernelUCB, ce qui est attendu considérant les performances de TS et UCB en bandits classiques. Cela montre qu'un certain écart demeure entre les approches performantes en pratique et celles validées en théorie. La porte est donc ouverte aux travaux futurs visant à fournir des garanties de convergence pour un Kernel TS *compétitif*.

Finalement, la dimensionnalité de l'espace des actions demeure un problème pour les approches de régression à noyau. D'un point de vue d'apprentissage en-ligne, le besoin d'acquérir un important nombre d'observations pour couvrir l'espace de recherche est problématique

parce qu'il repousse la convergence. La dimensionnalité n'est pas donc pas seulement une limite computationnelle de la régression à noyau, mais également un frein à l'optimisation en-ligne. La malédiction de la dimensionnalité a été abordée (Djolonga et al., 2013; Wang et al., 2016; Li et al., 2016) dans la régression GP sous l'hypothèse d'une dimensionnalité effective restreinte. Il serait pertinent de généraliser ces approches à la régression à noyau de régularisation adaptative de bruit inconnu.

# Chapitre 4

# Les bandits multi-objectifs

L'optimisation multi-objectif (Coello et al., 2007) est un sujet d'une grande importance pour les applications du monde réel. En effet, les problèmes d'optimisation sont souvent caractérisés par un certain nombre de mesures de performance incompatibles, voire contradictoires, nécessaires pour la tâche à accomplir. Par exemple, lors de la prise de décision sur les soins de santé à attribuer pour un patient atteint d'une maladie donnée, un compromis doit être effectué entre l'efficacité d'une procédure à traiter la maladie, les effets secondaires du traitement ainsi que son coût. Les problèmes d'optimisation multi-objectif sont souvent abordés en combinant les objectifs en une seule mesure (aussi connu sous le nom de *scalarisation*). De telles approches sont dites *a priori*, puisque la préférence sur les objectifs est articulée avant d'effectuer l'optimisation elle-même. Le défi réside donc dans la détermination de la fonction de préférence appropriée et de son paramétrage. Une autre façon de procéder à l'optimisation de multiples objectifs consiste à apprendre les compromis optimaux (le soi-disant ensemble Pareto-optimal (Ehrgott, 2012)). Une fois l'optimisation terminée, des techniques du domaine de la décision multi-critère sont utilisées pour aider un preneur de décision à sélectionner une solution finale dans l'ensemble Pareto-optimal. Ces techniques *a posteriori* peuvent nécessiter un grand nombre d'évaluations afin d'obtenir une estimation fiable des valeurs des objectifs pour toutes les solutions possibles. En effet, l'ensemble Pareto-optimal peut être assez grand, englobant la majorité, sinon la totalité, des solutions potentielles.

Dans ce chapitre, nous abordons le problème d'optimisation multi-objectif où la fonction de préférence *existe* a priori, mais peut être inconnue, auquel cas un utilisateur (expert) agit comme une boîte noire pour articuler les préférences. L'intégration de l'expert à la boucle d'apprentissage permet à ce dernier de fournir une rétroaction en sélectionnant son choix préféré étant donné un ensemble d'options - la fonction de préférence résidant dans sa tête. Plus précisément, nous considérons les problèmes où les observations sont stochastiques et coûteuses à obtenir (par exemple, impliquant un être humain dans la boucle). Le défi consiste donc à découvrir les meilleures solutions en fonction des observations aléatoires échantillonnées à partir de différentes distributions de densités inconnues. Nous formulons ce problème comme

des bandits multi-objectifs, où nous cherchons à trouver la solution qui maximise la fonction de préférence tout en maximisant la performance des solutions essayées lors de l'optimisation. À cet effet, nous proposons d'utiliser une approche de Thompson *sampling* (TS) adaptée aux bandits multi-objectifs : TS-MVN.

Considérons que le *bon choix* dénote l'option maximisant la fonction de préférence, c'est-à-dire l'option que l'expert choisirait s'il connaissait l'ensemble Pareto-optimal. Un algorithme d'apprentissage pour les bandits multi-objectifs vise à apprendre des estimations suffisamment précises des options disponibles pour permettre à l'expert d'effectuer de bons choix. Sa performance dépend donc de la robustesse de la fonction de préférence à la qualité des estimations. Ainsi, nous avons besoin d'une mesure caractérisant la qualité des estimations requises pour que l'option maximisant la fonction de préférence demeure inchangée. À cet effet, nous introduisons le concept de rayon de préférence. Ce dernier définit la plage de tolérance sur l'estimation des valeurs des objectifs de sorte que la préférence de l'expert demeure la même que si l'ensemble Pareto-optimal était connu. Nous utilisons ce concept pour fournir une analyse théorique de TS-MVN.

Finalement, nous réalisons des expériences empiriques qui supportent les résultats théoriques. Les résultats expérimentaux soulignent également l'importance de traiter les problèmes de bandits multi-objectifs comme tels au lieu de les scalariser pour les aborder sous forme de problèmes de bandits traditionnels. La convergence de TS-MVN, possédant des garanties de convergence, est ensuite comparée à une version alternative *complètement empirique*, mais pour laquelle aucune garantie n'est offerte. Enfin, des expériences préliminaires sont réalisées pour évaluer la possibilité de remplacer certaines rétroactions de l'expert par des fonctions automatiques puisque le fait d'avoir un expert dans le processus (pour articuler la fonction de préférence) pourrait être considéré comme une contrainte non négligeable. Les résultats montrent le potentiel de ces approches et proposent des pistes de recherches subséquentes.

## 4.1 Formulation du problème

Le problème de bandits multi-objectifs est décrit par un ensemble (fini, discret) d'actions $\mathcal{A}$, l'*espace de design* (*design space*), chacune étant associée à une observation $d$-dimensionnelle attendue et inconnue $\boldsymbol{\mu}_a := (\mu_{a,1}, \ldots, \mu_{a,d}) \in \mathcal{X} \subset \mathbb{R}^d$. Pour la simplicité, nous supposons un l'*espace des objectifs* (*objective space*) borné, soit $\mathcal{X} := [0,1]^d$. Dans ce jeu itératif, un agent interagit avec un environnement caractérisé par une *fonction de préférence* $f$. À chaque épisode $t \in \mathbb{N}_{>0}$, l'agent choisit une action $a_t$ à effectuer et obtient une observation bruitée $\boldsymbol{y}_t := (y_{t,1}, \ldots, y_{t,d})$ [1]. Il est important de noter que $\boldsymbol{y}_t$ est une observation bruitée de $\boldsymbol{\mu}_{a_t}$.

1. Dans ce chapitre, les vecteurs sont écrits en gras. Les opérateurs $+$, $-$, $\times$ et $\div$ appliqués entre un vecteur $\boldsymbol{v} = (v_1, \ldots, v_d)$ et un scalaire $s$ correspondent à l'opération entre chaque élément de $\boldsymbol{v}$ et $s$. Par exemple, $\boldsymbol{v}+s = (v_1 + s, \ldots, v_d + s)$. Ces mêmes opérateurs appliqués entre deux vecteurs $\boldsymbol{v} = (v_1, \ldots, v_d)$ et $\boldsymbol{u} = (u_1, \ldots, u_d)$ correspondent aux opérations par éléments de $\boldsymbol{v}$ et $\boldsymbol{u}$. Par exemple, $\boldsymbol{v} + \boldsymbol{u} = (v_1 + u_1, \ldots, v_d + u_d)$.

Soit l'ensemble des actions optimales $\mathcal{O} := \arg\max_{a \in \mathcal{A}} f(\boldsymbol{\mu}_a)$ et soit $\star \in \mathcal{O}$ une action optimale [2]. L'écart de sous-optimalité de l'action $a$, $\Delta_a := f(\boldsymbol{\mu}_\star) - f(\boldsymbol{\mu}_a)$, mesure la perte attendue lorsque l'action $a$ est sélectionnée à la place de l'action optimale. Le but de l'agent est de minimiser le pseudo-regret cumulatif attendu [3] :

$$\mathbb{E}[\mathfrak{R}(T)] \stackrel{\text{def}}{=} \mathbb{E}\Big[\sum_{t=1}^{T} \big(f(\boldsymbol{\mu}_\star) - f(\boldsymbol{\mu}_{a_t})\big)\Big] = \sum_{a \in \mathcal{A}} \sum_{t=1}^{T} \mathbb{P}[a_t = a] \Delta_a. \tag{4.1}$$

Cette quantité mesure la performance attendue d'un algorithme comparée à la performance attendue d'un algorithme optimal qui connaîtrait les distributions générant les observations, c'est-à-dire qui échantillonnerait toujours de la distribution avec l'espérance maximisant $f$.

**Remarque 17.** *Le problème de bandits traditionnel correspond à une instance spécifique du problème de bandits multi-objectifs décrit ici dans lequel $d = 1$ et $f(x) = x$.*

Si la fonction de préférence était connue, le problème se résumerait à un problème d'optimisation convexe (Agarwal et al., 2011). Cependant, dans de nombreuses situations, l'environnement fournissant la fonction de préférence est une personne, appelons-la l'*expert*. Malheureusement, les gens sont généralement incapables de scalariser leurs choix et préférences. Par conséquent, ils ne peuvent pas fournir explicitement leur fonction de préférence. Toutefois, compte tenu de plusieurs options, les experts peuvent indiquer celle(s) qu'ils préfèrent et peuvent donc être utilisés comme une boîte noire pour fournir des rétroactions dans la boucle d'apprentissage. Nous abordons ici le problème dans lequel la fonction $f$ n'est pas accessible, c'est-à-dire que l'on ne peut pas obtenir l'évaluation d'un point donné $\boldsymbol{x}$, soit $f(\boldsymbol{x})$. Les décisions reposent alors complètement sur l'évaluation d'options présentées à un expert.

Un algorithme de bandits multi-objectifs est une méthode (possiblement randomisée) pour sélectionner la prochaine action à entreprendre étant donné l'historique des choix antérieurs et des observations obtenues, $\mathcal{H}_t := \{a_s, \boldsymbol{y}_s\}_{s=1}^{t}$. Typiquement, il est supposé que l'algorithme propose une estimation $\boldsymbol{\theta}_{a,t}$ de $\boldsymbol{\mu}_a$ pour chaque action $a$ au temps $t$. Soit $\mathcal{O}_t := \arg\max_{a \in \mathcal{A}} f(\boldsymbol{\theta}_{a,t})$ l'ensemble des actions dont les estimations maximisent $f$. L'algorithme doit effectuer un compromis entre l'essai d'une action $a_t \in \mathcal{O}_t$, qui maximise potentiellement $f$, et l'acquisition d'une observation supplémentaire pour une action relativement méconnue dans le but d'améliorer son estimation. L'algorithme 8 décrit ce problème de bandits multi-objectifs.

**Optimalité de Pareto** Soient deux options $\boldsymbol{x} = (x_1, \ldots, x_d)$ et $\boldsymbol{y} = (y_1, \ldots, y_d)$ à $d$ dimensions. Il est considéré que $\boldsymbol{x}$ domine, ou Pareto-domine (Ehrgott, 2012), $\boldsymbol{y}$ (dénoté $\boldsymbol{x} \succeq \boldsymbol{y}$) si et seulement si $x_i > y_i$ pour au moins un $i$ et $x_i \geqslant y_i$ sinon. La relation de domination

2. La supposition d'une action optimale unique est effectuée sans perte de généralité puisque la présence d'actions optimales additionnelles ne peut que réduire la perte attendue (Agrawal and Goyal, 2012).

3. Aussi connu comme le regret scalarisé (*scalarized regret*) (Drugan and Nowe, 2013).

**Algorithme 8** Problème de bandits multi-objectifs

L'agent possède un estimateur trivial de $\boldsymbol{\mu}_a$ pour chaque action $a$.
À chaque épisode $t \geqslant 1$ :

1. L'agent présente des estimations $(\boldsymbol{\theta}_{a,t})_{a \in \mathcal{A}}$ de $(\boldsymbol{\mu}_a)_{a \in \mathcal{A}}$.
2. L'agent observe $\mathcal{O}_t$.
3. L'agent sélectionne une action $a_t$ étant donné $\mathcal{O}_t$.
4. L'agent observe $\boldsymbol{y}_t = \boldsymbol{\mu}_{a_t} + \boldsymbol{\xi}_t$, où $\boldsymbol{\xi}_t$ est un vecteur de variables aléatoires i.i.d.

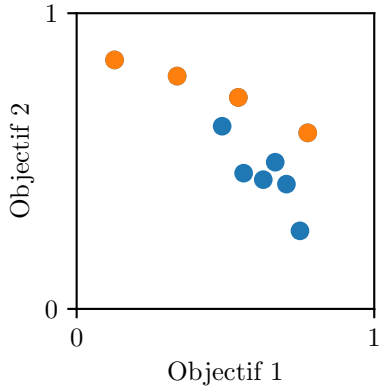



FIGURE 4.1 – Exemple d'options dominées (bleu) et non dominées (orange).

est stricte (dénotée $\boldsymbol{x} \succ \boldsymbol{y}$) si et seulement si $x_i > y_i$ pour tout $i = 1, \ldots, d$. Finalement, les deux vecteurs sont considérés incomparables (dénoté $\boldsymbol{x} \parallel \boldsymbol{y}$) si $\boldsymbol{x} \not\succeq \boldsymbol{y}$ et $\boldsymbol{y} \not\succeq \boldsymbol{x}$. On dit d'une option qu'elle est *Pareto-optimale* si elle n'est dominée par aucune autre option. Les options Pareteo-optimales représentent donc les meilleurs compromis entre les objectives et sont les seules options qui doivent être considérées dans une application. Ces options constituent le front de Pareto $\mathcal{P} = \{a \in \mathcal{A} : \nexists \boldsymbol{\mu}_b \succeq \boldsymbol{\mu}_a \quad \forall b \in \mathcal{A}\}$. La figure 4.1 montre un exemple d'options dominées et non dominées dans un espace à $d = 2$ objectifs. Un expert faisant face à un problème de décision multi-critères doit choisir son option non dominée préférée. Les options dominées sont évidemment éliminées par défaut.

## 4.2 Littérature

Le problème des bandits multi-objectifs a posteriori, où l'objectif est de découvrir l'ensemble du front de Pareto pour une prise de décision a posteriori, a déjà été abordé dans la littérature (Drugan and Nowe, 2013; Durand et al., 2014; Yahyaa et al., 2015). Ce problème est différent du problème d'optimisation a priori abordé ici. Le but des algorithmes dans le contexte a posteriori est de minimiser simultanément le regret de Pareto (*Pareto-regret*) et la métrique d'iniquité (*unfairness*). Aussi connue sous le nom d'$\varepsilon$-distance (Laumanns et al., 2002), le regret de Pareto associé à l'action $a$ est la valeur minimale $\varepsilon_a$ telle que $\boldsymbol{\mu}_a + \varepsilon_a$ n'est dominée par aucune autre action. En d'autres termes, toute action sur le front est considérée

comme également bonne par l'expert. C'est l'équivalent de considérer que $\mathcal{O} = \mathcal{P}$, ce qui correspond à la fonction de préférence

$$f(\boldsymbol{\mu}_\star) = 1 \quad \text{et} \quad f(\boldsymbol{\mu}_a) = 1 - \varepsilon_a,$$

telle que $\Delta_a = \varepsilon_a$. Notons que n'importe quel algorithme de bandits mono-objectifs optimisant un seul objectif pourrait minimiser le regret de Pareto sans égard aux autres objectifs. Ce comportement est contrôlé par la métrique d'iniquité mesurant la disparité entre le nombre d'allocations attribuées aux différentes actions formant le front de Pareto. L'idée est donc de forcer les algorithmes de bandits multi-objectifs a posteriori à explorer uniformément l'entièreté du front.

En dehors du domaine des bandits, les problèmes d'optimisation multi-objectifs (Zuluaga et al., 2013) visent à identifier l'ensemble Pareto-optimal $\mathcal{P}$ sans évaluer toutes les actions. La qualité d'une solution $\mathcal{S}$ est typiquement donnée par l'erreur d'hypervolume $V(\mathcal{P}) - V(\mathcal{S})$, où l'hypervolume $V(\mathcal{P})$ dénote le volume clôturé par l'origine et $\{\boldsymbol{\mu}_a\}_{a\in\mathcal{P}}$ (similairement pour $\mathcal{S}$). Cependant, l'erreur d'hypervolume ne fournit pas d'information à propos de la qualité des estimations des actions. L'identification seule du front de Pareto ne garantit pas que les actions sont bien estimées et donc qu'un expert serait en mesure d'effectuer le bon choix en se basant sur ces estimations.

Dans le problème des bandits multi-objectifs tel qu'introduit dans ce chapitre, la dynamique d'interaction avec l'expert indiquant sa préférence par le biais de l'ensemble $\mathcal{O}_t$ rappelle également les *dueling* bandits (Yue et al., 2009, 2012). Dans cette variante des bandits traditionnels, l'algorithme n'observe pas directement des récompenses associées aux actions, mais plutôt des ordonnancements de préférences par paires d'éléments fournies par un utilisateur (ici, l'expert). Les comparaisons sont effectuées entre des paires de récompenses obtenues avec les différentes actions. Ici, les comparaisons sont plutôt effectuées entre des options générées par un algorithme. Cette façon de procéder pour contrer l'inaccessibilité ou la difficulté de définir proprement une fonction de préférence a également été considérée récemment en *reinforcement learning* (RL) profond (*deep* RL) (Christiano et al., 2017).

## 4.3 Rayon de préférence

Dans le but de quantifier la qualité minimale requise des estimations pour permettre à un expert d'effectuer le bon choix, c'est-à-dire de prendre la même décision que s'il avait eu accès aux valeurs inconnues $\boldsymbol{\mu}_a$ pour tout $a \in \mathcal{A}$, nous proposons le concept de rayon de préférence. Soit l'estimation $\boldsymbol{\theta}_{a,t}$ associée à l'action $a$ à l'épisode $t$. Soit le front de Pareto estimé $\mathcal{P}_t := \{a \in \mathcal{A} : \nexists \boldsymbol{\theta}_{b,t} \succeq \boldsymbol{\theta}_{a,t} \quad \forall b \in \mathcal{A}\}$ étant donné ces options. Rappelons que, par définition, les options optimales sont données par $\mathcal{O}_t \subseteq \mathcal{P}_t$. Soit

$$B(\boldsymbol{c}, r) \subseteq \{\boldsymbol{x} \in \mathcal{X} : |x_i - c_i| \leqslant r,\ i = 1, \ldots, d\}$$

une boule de centre $\boldsymbol{c}$ et de rayon $r$. Nous définissons la quantité suivante, dépendante du problème, pour caractériser la difficulté d'un environnement de bandits multi-objectifs.

**Définition 6.** *Nous définissons une collection de rayons de préférences $(\rho_a)_{a\in\mathcal{A}}$ telle que si $\boldsymbol{\theta}_{a,t} \in B(\boldsymbol{\mu}_a, \rho_a)$ simultanément pour tout $a \in \mathcal{A}$, alors $\mathcal{O}_t \subseteq \mathcal{O}$.*

Autrement dit, si l'estimation de chaque action est dans la boule associée à cette action, alors les options préférées sont réellement associées à des actions optimales. Les rayons de préférence fournissent de l'information à la fois sur la rigidité de la fonction de préférence et sur la tolérance aux actions sous-optimales. Ces rayons correspondent à la *robustesse* de la fonction de préférence, c'est-à-dire à la quantité de variation simultanée pouvant être tolérée sur chaque action avant que l'ensemble des options optimales ne change. Le rayon $\rho_a$ est directement lié à l'écart $\Delta_a = f(\boldsymbol{\mu}_\star) - f(\boldsymbol{\mu}_a)$. Pour une action sous-optimale, un rayon large indique que cette action est loin d'être optimale. De plus, les rayons de préférence des actions sous-optimales dépendent du rayon de l'action optimale. Un rayon plus large pour l'action optimale implique des rayons plus petits pour les actions sous-optimales. Notons que si toutes les estimations sont contenues dans leur boule de préférence respective, alors la sélection de $a_t \in \mathcal{O}_t$ implique nécessairement un regret nul au temps $t$.

Soient les *poids* $\alpha_1, \ldots, \alpha_d \in [0, 1]$ tels que $\sum_{i=1}^d \alpha_i = 1$. La métrique $L_p$ pondérée

$$f(\boldsymbol{x}) = \left( \sum_{i=1}^{d} \alpha_i x_i^p \right)^{1/p}$$

avec $p \geqslant 1$ est souvent utilisée pour représenter des fonctions de décision. Cette fonction est connue sous le terme de scalarization linéaire quand $p = 1$ et comme la scalarization Chebyshev quand $p = \infty$. Les exemples suivants montrent le lien entre les rayons de préférence et l'écart de sous-optimalité pour ces deux fonctions populaires.

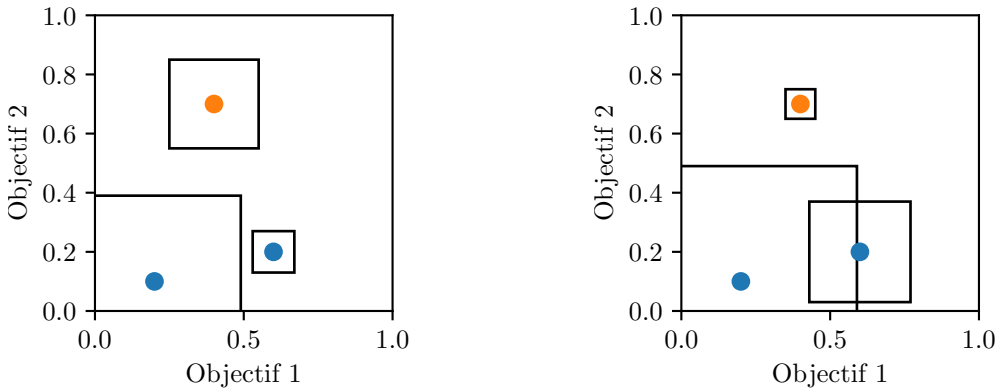
**Exemple 1** (Scalarisation linéaire)**.** *La fonction de scalarisation linéaire est donnée par*

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \alpha_i x_i.$$

*Considérons une action optimale $\star$ et une action sous-optimale $a$. Par définition des rayons de préférence, nous avons*

$$\begin{aligned}
\min_{\boldsymbol{x}_\star \in B(\boldsymbol{\mu}_\star, \rho_\star)} f(\boldsymbol{x}_\star) &> \max_{\boldsymbol{x}_a \in B(\boldsymbol{\mu}_a, \rho_a)} f(\boldsymbol{x}_a) \\
\sum_{i=1}^{d} (\alpha_i \mu_{\star,i} - \alpha_i \rho_\star) &> \sum_{i=1}^{d} (\alpha_i \mu_{a,i} + \alpha_i \rho_a) \\
f(\boldsymbol{\mu}_\star) - \rho_\star &> f(\boldsymbol{\mu}_a) + \rho_a \\
\Delta_a &> \rho_\star + \rho_a.
\end{aligned}$$

*La figure 4.2 montre des exemples de rayons de préférences pour une fonction linéaire.*




(a) $\rho_\star = 0.15$, $\rho_a = 0.07$, $\rho_b = 0.29$ (b) $\rho_\star = 0.05$, $\rho_a = 0.17$, $\rho_b = 0.39$

FIGURE 4.2 – Exemples de rayons de préférence autour de l'action optimale (orange) et des actions sous-optimales (bleu) pour la fonction de préférence linéaire $f(\boldsymbol{x}) = 0.4x_1 + 0.6x_2$.

**Exemple 2** (Scalarisation Chebyshev)**.** *La fonction de scalarisation Chebyshev (Bowman Jr, 1976) est donnée par*

$$f(\boldsymbol{x}) = \max_{1 \leqslant i \leqslant d} \alpha_i x_i.$$

*Considérons une action optimale $\star$ et sous-optimale $a$, et soient*

$$i_\star = \arg\max_{1 \leqslant i \leqslant d} \alpha_i(\mu_{\star,i} - \rho_\star), \quad i_a = \arg\max_{1 \leqslant i \leqslant d} \alpha_i(\mu_{a,i} + \rho_a).$$

*Par définition des rayons de préférence, nous avons*

$$\begin{aligned}
\min_{\boldsymbol{x}\star \in B(\boldsymbol{\mu}_\star, \rho_\star)} f(\boldsymbol{x}_\star) &> \max_{\boldsymbol{x}_a \in B(\boldsymbol{\mu}_a, \rho_a)} f(\boldsymbol{x}_a) \\
\max_{1 \leqslant i \leqslant d} \alpha_i(\mu_{\star,i} - \rho_\star) &> \max_{1 \leqslant i \leqslant d} \alpha_i(\mu_{a,i} + \rho_a) \\
\alpha_{i_\star}\mu_{\star,i_\star} - \alpha_{i_\star}\rho_\star &> \alpha_{i_a}\mu_{a,i_a} + \alpha_{i_a}\rho_a \\
f(\boldsymbol{\mu}_\star) - \alpha_{i_\star}\rho_\star &> f(\boldsymbol{\mu}_a) + \alpha_{i_a}\rho_a \\
\Delta_a &> \alpha_{i_\star}\rho_\star + \alpha_{i_a}\rho_a.
\end{aligned}$$

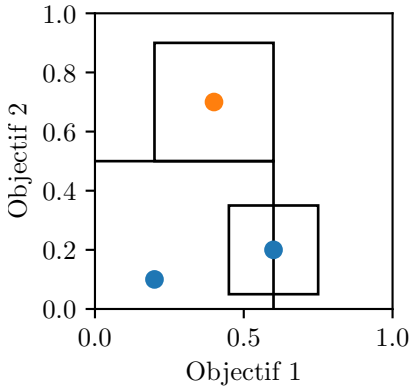
*La difficulté ici est que $i_\star$ et $i_a$ dépendent respectivement de $\rho_\star$ et $\rho_a$. Considérons un environnement à deux objectifs, nous pouvons définir les seuils*

$$\tau_\star = \frac{\alpha_2\mu_{\star,2} - \alpha_1\mu_{\star,1}}{\alpha_2 - \alpha_1}, \quad \tau_a = \frac{\alpha_1\mu_{a,1} - \alpha_2\mu_{a,2}}{\alpha_2 - \alpha_1}$$

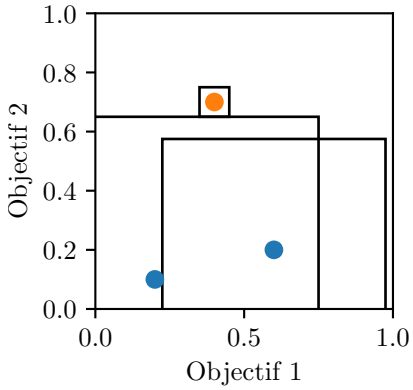
*tels que*

$$i_\star = \begin{cases} 1 & si \quad \rho_\star > \tau_\star \\ 2 & sinon \end{cases}, \quad i_a = \begin{cases} 1 & si \quad \rho_a < \tau_a \\ 2 & sinon. \end{cases}$$

*La figure 4.3 montre des exemples de rayons de préférence avec une fonction Chebyshev.*




(a) $\rho_\star = 0.2$, $\rho_a = 0.15$, $\rho_b = 0.4$




(b) $\rho_\star = 0.05$, $\rho_a = 0.375$, $\rho_b = 0.55$

FIGURE 4.3 – Exemples de rayons de préférence autour de l'action optimale (orange) et des actions sous-optimales (bleu) pour une fonction de Chebyshev avec $\alpha_1 = 0.4$ et $\alpha_2 = 0.6$.

Outre les métriques $L_p$, on retrouve également les fonctions de scalarization basées sur des contraintes. Par exemple, avec la technique de scalarisation $\varepsilon$-contrainte, l'expert attribue une contrainte à tous les objectifs à l'exception d'un objectif *cible* $\ell$. Toutes les options qui ne respectent pas l'une des contraintes obtiennent une valeur de 0, tandis que celles qui les respectent toutes obtiennent une valeur de $x_\ell$. L'exemple suivant montre la relation entre le rayon de préférence et l'écart de sous-optimalité étant donné une fonction de préférence articulée avec une approche de type $\varepsilon$-contrainte.

**Exemple 3** (Epsilon-contrainte)**.** *Une fonction $\varepsilon$-contrainte est donnée par*

$$f(\boldsymbol{x}) = \begin{cases} x_\ell & si \quad x_i \geqslant \varepsilon_i \quad \forall i \in \{1, \dots, d\}, i \neq \ell \\ 0 & sinon. \end{cases}$$

*Considérons une action optimale $\star$ et sous-optimale $a$. Par définition des rayons de préférence, nous avons*

$$\rho_\star \leqslant \min_{1 \leqslant i \leqslant d, i \neq \ell} \mu_{\star,i} - \varepsilon_i.$$

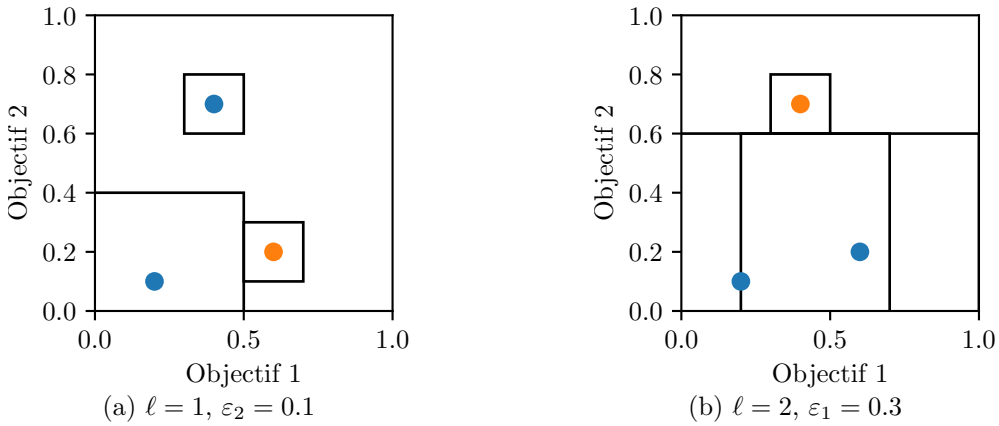
*Nous décomposons $\rho_a = \underline{\rho}_a + \bar{\rho}_a$ de telle sorte que*

$$\underline{\rho}_a = \min\{0, \max_{1 \leqslant i \leqslant d, i \neq \ell} \varepsilon_i - \mu_{a,i}\}$$

*dénote le rayon requis pour que l'action $a$ respecte les contraintes, c'est-à-dire pour obtenir $f(\boldsymbol{\mu}_a) > 0$, et $\bar{\rho}_a$ dénote l'extra conduisant à une réduction de l'écart de sous-optimalité. Finalement, nous obtenons*

$$\mu_{\star,\ell} - \rho_\star > \mu_{a,\ell} + \underline{\rho}_a + \bar{\rho}_a \quad et \quad \Delta_a > \rho_\star + \rho_a.$$

*La figure 4.4 montre des exemples de rayons de préférence pour des fonctions $\varepsilon$-contraintes.*




(a) $\ell = 1$, $\varepsilon_2 = 0.1$ (b) $\ell = 2$, $\varepsilon_1 = 0.3$

FIGURE 4.4 – Exemples de rayons de préférence autour de l'action optimale (orange) et des actions sous-optimales (bleu) pour deux configurations d'une fonction $\varepsilon$-contrainte.

## 4.4 TS-MVN

Maintenant que nous possédons une manière de quantifier la qualité des estimations, nous introduisons une méthode permettant de générer de telles estimations. L'algorithme TS (voir la section 1.1.2) maintient une distribution a posteriori $\pi_{a,t}$ sur la moyenne $\boldsymbol{\mu}_a$ étant donné un a priori et l'historique des observations $\mathcal{H}_{t-1}$. À l'épisode $t$, une option $\boldsymbol{\theta}_{a,t}$ est échantillonnée de chaque distribution postérieure $\pi_{a,t}$. L'algorithme choisit une action $a_t \in \mathcal{O}_t$. Rappelons que $\mathcal{O}_t := \arg\max_{a \in \mathcal{A}} f(\boldsymbol{\theta}_{a,t})$. Ainsi, $\mathbb{P}[a_t = a]$ est proportionnel à la probabilité a posteriori que $a$ maximise la fonction de préférence étant donné l'historique $\mathcal{H}_{t-1}$. Soit le nombre de tirages $N_{a,t} = \sum_{s=1}^{t} \mathbb{I}[a_s = a]$ de l'action $a$ jusqu'à l'épisode $t$ (inclusivement). La moyenne empirique est respectivement donnée par

$$\widehat{\boldsymbol{\mu}}_{a,t} = \frac{\sum_{s=1:a_s=a}^{t} \boldsymbol{y}_s}{N_{a,t}}. \tag{4.2}$$

Soit les a priori $\boldsymbol{\Sigma}_0$ et $\boldsymbol{\mu}_0$. Considérant un a priori normal multi-varié (MVN), le postérieur sur $\boldsymbol{\mu}_a$ conditionné sur les observations obtenues jusqu'au temps $t$ (inclusivement) est donné par la distribution $\mathcal{N}_d(\tilde{\boldsymbol{\mu}}_{a,t}, \tilde{\boldsymbol{\Sigma}}_{a,t})$, où

$$\tilde{\boldsymbol{\Sigma}}_{a,t} = \left(\boldsymbol{\Sigma}_0^{-1} + N_{a,t}\boldsymbol{\Sigma}_a^{-1}\right)^{-1} \quad \text{et} \quad \tilde{\boldsymbol{\mu}}_{a,t} = \tilde{\boldsymbol{\Sigma}}_{a,t}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + N_{a,t}\boldsymbol{\Sigma}_a^{-1}\widehat{\boldsymbol{\mu}}_{a,t})$$

pour une matrice de covariance $\boldsymbol{\Sigma}_a$ connue. Comme l'hypothèse de la connaissance de $\boldsymbol{\Sigma}_a$ peut se révéler irréaliste en pratique, une alternative consiste à considérer la matrice de covariance non informative $\boldsymbol{\Sigma}_a = \boldsymbol{I}_d$. Avec les prieurs [4] $\boldsymbol{\mu}_0 = \mathbf{0}_{d\times 1}$ et $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_d$, [5] l'algorithme correspond à une extension directe du TS uni-dimensionnel d'a priori normal (Agrawal and Goyal, 2013). L'algorithme 9 présente la procédure résultante de TS d'a priori MVN, TS-MVN.

4. Sachant que $\mathcal{X} = [0, 1]^d$.
5. $\mathbf{0}_{d\times 1}$ indique un vecteur colonne nul de $d$ éléments et $\boldsymbol{I}_d$ indique une matrice identité de dimension $d \times d$.

**Algorithme 9** TS-MVN

initialiser $\boldsymbol{\mu}_a = \mathbf{0}_d$ pour chaque $a \in \mathcal{A}$
**for all** épisode $t \geqslant 1$ **do**
  **for all** action $a \in \mathcal{A}$ **do**
    calculer $\tilde{\boldsymbol{\mu}}_{a,t-1} = \frac{N_{a,t-1}}{N_{a,t-1}+1}\widehat{\boldsymbol{\mu}}_{a,t-1}$ et $\tilde{\boldsymbol{\Sigma}}_{a,t-1} = \frac{1}{N_{a,t-1}+1}\boldsymbol{I}_d$
    échantillonner $\boldsymbol{\theta}_{a,t} \sim \mathcal{N}_d(\tilde{\boldsymbol{\mu}}_{a,t-1}, \tilde{\boldsymbol{\Sigma}}_{a,t-1})$
  **end for**
  observer $\mathcal{O}_t = \arg\max_{a\in\mathcal{A}} f(\boldsymbol{\theta}_{a,t})$
  jouer $a_t \in \mathcal{O}_t$, observer $\boldsymbol{y}_t$ et calculer $\widehat{\boldsymbol{\mu}}_{a,t}$
**end for**

La proposition suivante fournit des bornes de regret générales pour TS-MVN, indépendantes de la fonction de préférence. Le théorème suivant spécialise ces bornes de regret pour trois familles de fonctions de préférence bien connues en exploitant les relations entre les rayons de préférence et l'écart de sous-optimalité, comme discuté dans les exemples précédents.

**Proposition 1** (TS-MVN)**.** *Sous l'hypothèse de bruit $\sigma$-sous-gaussien avec $\sigma^2 \leqslant 1/(4d)$, le regret attendu TS-MVN (algorithme 9) est borné par*

$$\mathbb{E}[\mathfrak{R}(T)] \leqslant \sum_{\substack{a\in\mathcal{A}\\ a\neq\star}} \left[(C(d)+4d)(1+\sigma)\Delta_a \frac{\ln(dT\Delta_a^2)}{\rho_\star^2} + \frac{4}{\Delta_a} + 2\Delta_a \frac{\ln(dT\Delta_a^2)}{(\rho_a - r_a)^2} + 2\sigma^2\Delta_a \frac{\ln(dT\Delta_a^2)}{r_a^2}\right],$$

*où $\rho_\star$, $\rho_a$ sont des rayons de préférences, $r_a < \rho_a$ et $C(d)$ est tel que $e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}} \leqslant \frac{d}{i^2}$ pour $i \geqslant C(d)$ (voir la remarque 18). La procédure d'analyse est détaillée à la section suivante.*

**Théorème 3** (TS-MVN avec fonctions de préférences pré-définies)**.** *Sous l'hypothèse d'une fonction de préférence linéaire (exemple 1), Chebyshev (exemple 2) ou $\varepsilon$-contrainte (exemple 3), et de bruit $\sigma$-sous-gaussien avec $\sigma^2 \leqslant 1/(4d)$, le regret attendu de TS-MVN (algorithme 9) est borné par*

$$\mathbb{E}[\mathfrak{R}(T)] \leqslant \sum_{\substack{a\in\mathcal{A}\\ a\neq\star}} \left[(8C(d)+24d+18+72\sigma^2)(1+\sigma)^2 \frac{\ln(dT\Delta_a^2)}{\Delta_a} + \frac{4}{\Delta_a}\right],$$

*où $C(d)$ est tel que $e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}} \leqslant \frac{d}{i^2}$ pour $i \geqslant C(d)$ (voir la remarque 18). Cette borne sur le regret est donc d'ordre $\mathcal{O}(\sqrt{dNT}\ln d + \sqrt{dNT\ln N})$, où $N = |\mathcal{A}|$. Plus spécifiquement, pour $d \leqslant \ln N$, cette borne sur le regret est d'ordre $\mathcal{O}(\sqrt{dNT\ln N})$. La procédure d'analyse est détaillée à la section suivante.*

**Remarque 18.** *Pour $d = 1$, il est possible de prendre $C(d) = e^{14}$. Pour $d = 2$ et $d = 3$, il est possible de prendre respectivement $C(d) = e^{24}$ et $C(d) = e^{35}$, et ainsi de suite pour $d \in \mathbb{N}_{>0}$.*

**Convergence de TS-MVN** La dimension de l'espace des objectifs influence fortement la convergence de TS-MVN. Pour $d = 1$, l'ordre de la borne sur le regret donnée par le théorème 3 rencontre l'ordre de la borne sur le regret de TS-Normal dans un problème de bandits

mono-objectifs (Agrawal and Goyal, 2013) sous l'hypothèse d'observations bornées dans l'intervalle $[0, 1]$. Cependant, nous observons que la tolérance au bruit décroit linéairement avec la dimension $d$ de l'espace des objectifs. Cela signifie que plus l'espace des objectifs comporte de dimensions, moins grand doit être le bruit pour garantir que cette borne demeure valide, *étant donnée la présente analyse*.

Cela est justifié par la difficulté associée à l'estimation simultanée de plusieurs distributions. En mono-objectif (Agrawal and Goyal, 2013), il est considéré que l'action optimale est sélectionnée lorsque toutes les actions sont simultanément bien estimées et bien échantillonnées. En multi-objectifs, toutes les actions doivent être simultanément bien estimées *dans toutes leurs dimensions* et bien échantillonnées *dans toutes leurs dimensions* pour qu'une action optimale soit sélectionnée. L'intervalle de confiances autour d'un estimateur multi-dimensionnel (lemme 12) se resserre plus lentement en fonction du nombre de dimensions et la concentration des variables normales $d$-dimensionnelles (lemme 13) est plus difficile à contrôler. L'obtention de *conditions gagnantes* jointes sur l'ensemble des dimensions, simultanément pour chaque action, demeure donc un défi.

## 4.5 Analyse théorique

Cette section débute par la preuve de la proposition 1 fournissant une borne sur le regret de TS-MVN (algorithm 9), indépendante de la fonction de préférence. Les relations entre l'écart de sous-optimalité et les rayons de préférence pour trois familles de fonctions de préférence sont ensuite utilisées pour obtenir le théorème 3. Rappelons que TS-MVN sélectionne l'action $a_t = \arg\max_{a \in \mathcal{A}} \theta_{a,t}$, avec $\theta_{a,t} \sim \mathcal{N}_d(\tilde{\boldsymbol{\mu}}_{a,t-1}, \tilde{\boldsymbol{\Sigma}}_{a,t-1})$.

### 4.5.1 Preuve de la proposition 1

L'analyse suivante étend les travaux de Agrawal and Goyal (2013) du cas unidimensionnel au cas $d$-dimensionnel. L'équation 4.1 peut s'exprimer comme

$$\mathbb{E}[\mathfrak{R}(T)] = \sum_{a \in \mathcal{A}, a \neq \star} \Delta_a \sum_{t=1}^{T} \mathbb{P}[a_t = a],$$

où le contrôle du regret s'effectue à travers le contrôle de $\mathbb{P}[a_t = a]$. La preuve s'appuie sur plusieurs lemmes (voir l'annexe A), qui étendent les inégalités de Chernoff ainsi que les résultats de concentration et d'anti-concentration pour de variables gaussiennes du cas unidimensionnel au cas $d$-dimensionnel en utilisant les concepts de Pareto-domination et de rayons de préférence. Nous introduisons les quantités et événements suivants pour contrôler la qualité des distributions a posteriori ainsi que la qualité de leurs échantillons.

**Définition 7** (Quantités $r_a$)**.** *Pour chaque action sous-optimale $a$, nous choisissons une quantité $r_a < \rho_a$, où $\rho_a$ est un rayon de préférence. Par définition des rayons de préférences,*

*nous avons* $\boldsymbol{\mu}_a \prec \boldsymbol{\mu}_a + r_a \prec \boldsymbol{\mu}_a + \rho_a$. *Rappelons que* $f(\boldsymbol{x}) < f(\boldsymbol{x}')$ *si* $\boldsymbol{x} \prec \boldsymbol{x}'$. *Par conséquent nous avons* $f(\boldsymbol{\mu}_a) < f(\boldsymbol{\mu}_a + r_a) < f(\boldsymbol{\mu}_a + \rho_a) < f(\boldsymbol{\mu}_\star - \rho_\star)$.

**Définition 8** (Événements $E^{\mu}_{a,t}$, $E^{\theta}_{a,t}$)**.** *Pour chaque action sous-optimale* $a$*, nous définissons l'événement* $E^{\mu}_{a,t}$ *dénotant la situation* $\tilde{\boldsymbol{\mu}}_{a,t-1} \prec \boldsymbol{\mu}_a + r_a$ *et l'événement* $E^{\theta}_{a,t}$ *dénotant la situation où* $\boldsymbol{\theta}_{a,t} \prec \boldsymbol{\mu}_a + \rho_a$. *Plus précisément, ils représentent respectivement les situations où l'action sous-optimale* $a$ *est* bien *estimée et bien échantillonnée.*

**Définition 9** (Filtration $\mathcal{H}_t$)**.** *La filtration* $\mathcal{H}_t := \{a_s, \boldsymbol{y}_s\}_{s=1}^{t}$ *représente le passé.*

Nous décomposons ensuite, pour une action sous-optimale $a$,

$$\sum_{t=1}^{T} \mathbb{P}[a_t = a] = \underbrace{\sum_{t=1}^{T} \mathbb{P}[a_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t}]}_{(A)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}[a_t = a, E^{\mu}_{a,t}, \overline{E^{\theta}_{a,t}}]}_{(B)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}[a_t = a, \overline{E^{\mu}_{a,t}}]}_{(C)}$$

et contrôlons séparément chaque partie. La partie (A) traite la situation où $a$ est sélectionnée alors qu'elle est bien estimée et bien échantillonnée. Par définition des rayons de préférence, cela ne peut se produire si l'action optimale est bien échantillonnée. Ainsi, cette partie peut être contrôlée en limitant les mauvaises estimations et les mauvais échantillons pour l'action optimale. La partie (B) traite la situation où $a$ est sélectionnée tout en étant bien estimée, mais mal échantillonnée. Nous contrôlons cela en utilisant les inégalités de concentration gaussiennes. Enfin, la partie (C) traite de la situation où $a$ est sélectionnée alors qu'elle est mal estimée. Nous contrôlons cela en utilisant les inégalités de Chernoff. En réunissant les résultats suivants, nous obtenons la proposition 1.

### Contrôle de (A)

Par définition du TS, pour que l'action sous-optimale $a$ soit sélectionnée à l'épisode $t$, nous devons (au moins) avoir $f(\boldsymbol{\theta}_{a,t}) > f(\boldsymbol{\theta}_{\star,t})$. Par définition de l'événement $E^{\theta}_{a,t}$ et des rayons de préférence, nous avons $f(\boldsymbol{\theta}_{a,t}) < f(\boldsymbol{\theta}_{\star,t})$ si $\boldsymbol{\theta}_{\star,t} \succ \boldsymbol{\mu}_\star - \rho_\star$. Il s'agit donc de contrôler la probabilité que l'action optimale soit *mal* échantillonnée. Soit l'épisode $\tau_k$ auquel l'action $\star$ est sélectionnée par la $k^{\text{e}}$ fois, pour $k \geqslant 1$ et $\tau_0 = 0$. Notons que pour n'importe quelle action

$a$, $\tau_k > T$ pour $k > N_{a,t}$. Aussi, $\tau_T \geqslant T$. Ainsi nous avons

$$\begin{aligned}
(A) &= \sum_{t=1}^{T} \mathbb{P}[a_t = a, E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_{t-1}] \\
&\leqslant \sum_{t=1}^{T} \mathbb{P}[f(\boldsymbol{\theta}_{a,t}) > f(\boldsymbol{\theta}_{\star,t}), E^{\mu}_{a,t}, E^{\theta}_{a,t} | \mathcal{H}_{t-1}] \\
&\leqslant \sum_{t=1}^{T} \mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star} | \mathcal{H}_{t-1}] \\
&\leqslant \sum_{k=0}^{L} \mathbb{E}\Big[ \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star} | \mathcal{H}_{t-1}] \Big] + \sum_{t=\tau_L+1}^{T} \mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star}, N_{\star,t-1} \geqslant L | \mathcal{H}_{t-1}].
\end{aligned} \tag{4.3}$$

La deuxième inégalité utilise le fait que l'échantillonnage de $\boldsymbol{\theta}_{\star,t}$ est indépendant des événements $E^{\mu}_{a,t}$ et $E^{\theta}_{a,t}$. La dernière inégalité utilise l'observation que $\mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star} | \mathcal{H}_{t-1}]$ est fixe étant donné $\mathcal{H}_{t-1}$ et change seulement lorsque la distribution de $\boldsymbol{\theta}_{\star,t}$ change, c'est-à-dire seulement lorsque l'action $\star$ est sélectionnée. La première partie dénombre les épisodes requis avant que l'action $\star$ ne soit sélectionnée $L$ fois. La deuxième partie dénombre les épisodes où $\star$ est *mal* échantillonnée après été essayée $L$ fois. Nous contrôlons la première somme avec le lemme suivant.

**Lemme 7** (Basé sur le lemme 6 de Agrawal and Goyal (2013))**.** *Soit l'épisode $\tau_k$ où l'action $\star$ est sélectionnée pour la $k^e$ fois. Ainsi, pour tout $d \in \mathbb{N}_{>0}$ et $\sigma^2 \leqslant 1/(4d)$,*

$$\mathbb{E}\Big[ \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star} | \mathcal{H}_{t-1}] \Big] \leqslant C(d) + 4d,$$

*où $C(d)$ est tel que $e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i^d}}} \leqslant \frac{d}{i^2}$ pour $i \geqslant C(d)$.*

*Démonstration.* La démarche se base de près sur la preuve du lemme 6 d'Agrawal and Goyal (2013). Rappelons que $\tau_k$ dénote l'épisode où l'action $\star$ est sélectionnée pour la $k^{\text{e}}$ fois. Soit une variable aléatoire $\Theta_j$ distribuée selon une loi normale multivariée $\mathcal{N}_d(\tilde{\boldsymbol{\mu}}_{\star,\tau_j+1}, (\boldsymbol{I}_d + N_{\star,\tau_j+1}\boldsymbol{I}_d)^{-1})$. Soit une variable géométrique $G_j$ dénotant le nombre d'essais consécutifs indépendants avant d'obtenir $\Theta_j \succ \boldsymbol{\mu}_{\star} - \rho_{\star}$. Observons ensuite que

$$\mathbb{E}\Big[ \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_{\star} - \rho_{\star} | \mathcal{H}_{t-1}] \Big] \leqslant \mathbb{E}[G_j] = \sum_{i=1}^{\infty} \mathbb{P}[G_j \geqslant i].$$

Nous voulons borner la valeur attendue de $G_j$ par une constante pour tout $j$. Considérons n'importe quel entier $i \geqslant 1$, soit $z = \sqrt{\ln i^{1/d}}$ et soit $\text{MAX}_i$ la *préférence maximale* de $i$ échantillons indépendants $\Theta_j$, c'est-à-dire $\max_{1 \leqslant i \leqslant j} f(\Theta_j)$. Pour la suite, nous abrégeons $\tilde{\boldsymbol{\mu}}_{\star,\tau_j+1}$

comme $\tilde{\boldsymbol{\mu}}_\star$ et $N_{\star,\tau_j+1}$ comme $N_\star$. Ainsi,

$$\begin{aligned}\mathbb{P}[G_j < i] &\geqslant \mathbb{P}[\mathrm{MAX}_i \succ \boldsymbol{\mu}_\star - \rho_\star] \\ &\geqslant \mathbb{P}\Big[\mathrm{MAX}_i \succ \tilde{\boldsymbol{\mu}}_\star + \frac{z}{\sqrt{N_\star}} \Big| \tilde{\boldsymbol{\mu}}_\star + \frac{z}{\sqrt{N_\star}} \succeq \boldsymbol{\mu}_\star - \rho_\star\Big] \cdot \mathbb{P}\Big[\tilde{\boldsymbol{\mu}}_\star + \frac{z}{\sqrt{N_\star}} \succeq \boldsymbol{\mu}_\star - \rho_\star\Big].\end{aligned}$$

En utilisant les résultats d'anti-concentration pour une variable gaussienne $d$-dimensionnelle (lemme 14), nous obtenons

$$\begin{aligned}\mathbb{P}\Big[\mathrm{MAX}_i \succ \tilde{\boldsymbol{\mu}}_\star + \frac{z}{\sqrt{N_\star}} \Big| \tilde{\boldsymbol{\mu}}_\star + \frac{z}{\sqrt{N_\star}} \succeq \boldsymbol{\mu}_\star - \rho_\star\Big] &\geqslant 1 - \left(1 - \left(\frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-z^2/2}\right)^d\right)^i \\ &= 1 - \left(1 - \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln i^{1/d}}}{(\ln i^{1/d} + 1)} \frac{1}{\sqrt{i^{1/d}}}\right)^d\right)^i \\ &\geqslant 1 - \left(1 - \left(\frac{1}{\sqrt{18\pi d i^{1/d} \ln i}}\right)^d\right)^i \\ &\geqslant 1 - e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}},\end{aligned}$$

où la seconde inégalité utilise le fait que $\ln i^{1/d} + 1 < 3 \ln i$ pour $d \geqslant 1$ et la dernière inégalité utilise le fait que $1 - x < e^{-x}$ pour $x < 1$. Puis, en utilisant les bornes de Chernoff $d$-dimensionnelles (lemme 12), nous obtenons

$$\mathbb{P}[\tilde{\boldsymbol{\mu}}_\star \succeq \boldsymbol{\mu}_\star - \frac{z}{\sqrt{N_\star}}] \geqslant 1 - d e^{-\frac{z^2}{2\sigma^2}} = 1 - \frac{d}{i^{1/(2d\sigma^2)}}.$$

En substituant, nous obtenons

$$\mathbb{P}[G_j < i] \geqslant \left(1 - e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}}\right) \cdot \left(1 - \frac{d}{i^{1/(2d\sigma^2)}}\right) \geqslant 1 - \frac{d}{i^{1/(2d\sigma^2)}} - e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}}$$

et

$$\mathbb{E}[G_j] = \sum_{i \geqslant 1} (1 - \mathbb{P}[G_j < i]) \leqslant \sum_{i \geqslant 1} \left(\frac{d}{i^{1/(2d\sigma^2)}} + e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}}\right) \leqslant C(d) + 2d \sum_{i \geqslant 1} \frac{1}{i^{1/(2d\sigma^2)}},$$

où $C(d)$ est tel que $e^{-\frac{\sqrt{i}}{\sqrt{18\pi d \ln i}^d}} \leqslant \frac{d}{i^{1/(2d\sigma^2)}}$ pour $i \geqslant C(d)$. Nous remarquons que $\sigma^2 \leqslant 1/(4d)$ est nécessaire pour que la somme converge. □

Nous contrôlons maintenant la deuxième somme de l'équation 4.3 en contrôlant la probabilité d'échantillonner une *mauvaise* valeur de $\boldsymbol{\theta}_{\star,t}$ quand $N_{\star,t} > L$. Soit l'événement $E_{\star,t}$ dénotant

la situation où $\tilde{\boldsymbol{\mu}}_{\star,t-1} \succ \boldsymbol{\mu} - \sigma\rho_\star/(1+\sigma)$. Nous avons alors

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_\star - \rho_\star, N_{\star,t-1} > L|\mathcal{H}_{t-1}] \\
&\quad\leqslant \mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \tilde{\boldsymbol{\mu}}_{\star,t-1} - \frac{\rho_\star}{1+\sigma}, E_{\star,t}, N_{\star,t-1} > L|\mathcal{H}_{t-1}] + \mathbb{P}[\overline{E_{\star,t}}, N_{\star,t-1} > L|\mathcal{H}_{t-1}] \\
&\quad\leqslant \mathbb{P}[\boldsymbol{\theta}_{\star,t} \notin B\Big(\tilde{\boldsymbol{\mu}}_{\star,t-1}, \frac{\rho_\star}{1+\sigma}\Big), E_{\star,t}, N_{\star,t-1} > L|\mathcal{H}_{t-1}] \\
&\qquad + \mathbb{P}[\tilde{\boldsymbol{\mu}}_{\star,t-1} \notin B\Big(\boldsymbol{\mu} - \frac{\sigma\rho_\star}{1+\sigma}\Big), N_{\star,t-1} > L|\mathcal{H}_{t-1}] \\
&\quad\leqslant \frac{d}{2} e^{-\frac{L\rho_\star^2}{2(1+\sigma)^2}} + 2de^{-\frac{L\rho_\star^2}{2(1+\sigma)^2}},
\end{aligned}$$

où la dernière inégalité utilise les lemmes 12 et 13. En utilisant $L = 2(1+\sigma)^2 \frac{\ln(dT\Delta_a^2)}{\rho_\star^2}$, nous obtenons

$$\mathbb{P}[\boldsymbol{\theta}_{\star,t} \not\succ \boldsymbol{\mu}_\star - \rho_\star, N_{\star,t-1} > L|\mathcal{H}_{t-1}] \leqslant \frac{5}{2T\Delta_a^2}. \tag{4.4}$$

Finalement, nous utilisons le lemme 7 ainsi que l'équation 4.4 dans l'équation 4.3 pour obtenir

$$(A) \leqslant (2C(d) + 8d)(1+\sigma)^2 \frac{\ln(dT\Delta_a^2)}{\rho_\star^2} + \frac{5}{2\Delta_a^2}$$

avec $\sigma^2 \leqslant 1/(4d)$, où $C(d)$ est tel que $e^{\frac{\sqrt{i}}{\sqrt{18\pi d \ln i^d}}} \leqslant \frac{d}{i^2}$ pour $i \geqslant C(d)$.

**Contrôle de (B)**

Il s'agit ici de contrôler la probabilité de *mal* échantillonner l'action sous-optimale $a$ sachant qu'elle a été essayée au moins $L$ fois. Rappelons que la filtration $\mathcal{H}_{t-1}$ est telle que l'événement $E_{a,t}^\mu$ se produit. Ensuite, nous décomposons

$$\begin{aligned}
(B) &= \sum_{t=1}^T \mathbb{P}[a_t = a, \overline{E_{a,t}^\theta}, E_{a,t}^\mu, N_{a,t-1} \leqslant L|\mathcal{H}_{t-1}] + \sum_{t=1}^T \mathbb{P}[a_t = a, \overline{E_{a,t}^\theta}, E_{a,t}^\mu, N_{a,t-1} > L|\mathcal{H}_{t-1}] \\
&\leqslant \mathbb{E}\Big[\sum_{t=1}^T \mathbb{I}[a_t = a, N_{a,t-1} \leqslant L|\mathcal{H}_{t-1}]\Big] + \sum_{t=1}^T \mathbb{P}[\boldsymbol{\theta}_{a,t} \not\prec \boldsymbol{\mu}_a + \rho_a, N_{a,t-1} > L|\mathcal{H}_{t-1}] \\
&\leqslant L + \sum_{t=1}^T \mathbb{P}[\boldsymbol{\theta}_{a,t} \not\prec \tilde{\boldsymbol{\mu}}_{a,t-1} + (\rho_a - r_a), N_{a,t-1} > L|\mathcal{H}_{t-1}] \\
&\leqslant L + T\frac{d}{2} e^{-\frac{L(\rho_a - r_a)^2}{2}}.
\end{aligned}$$

La première inégalité utilise l'observation que $\mathbb{P}[a_t = a|\mathcal{H}_{t-1}]$ est fixe étant donné $\mathcal{H}_{t-1}$ ainsi que la définition de l'événement $\overline{E_{a,t}^\theta}$. La deuxième inégalité utilise le fait que l'événement $E_{a,t}^\mu$ se produit. La dernière inégalité utilise le lemme 13. En utilisant $L = 2\frac{\ln(dT\Delta_a^2)}{(\rho_a - r_a)^2}$, nous obtenons

$$(B) \leqslant 2\frac{\ln(dT\Delta_a^2)}{(\rho_a - r_a)^2} + \frac{1}{2\Delta_a^2}.$$

**Contrôle de (C)**

Similairement à ce qui a été effectué pour (B), il s'agit ici de contrôler la probabilité de *mal* estimer l'action sous-optimale $a$ sachant qu'elle a été essayée au moins $L$ fois. Nous décomposons donc

$$\begin{aligned}(C) &\leqslant \sum_{t=1}^{T} \mathbb{P}[a_t = a, \overline{E^{\mu}_{a,t}}, N_{a,t-1} \leqslant L | \mathcal{H}_{t-1}] + \sum_{t=1}^{T} \mathbb{P}[a_t = a, \overline{E^{\mu}_{a,t}}, N_{a,t-1} \geqslant L | \mathcal{H}_{t-1}] \\ &\leqslant \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}[a_t = a, N_{a,t-1} \leqslant L | \mathcal{H}_{t-1}]\Big] + \sum_{t=1}^{T} \mathbb{P}[\overline{E^{\mu}_{a,t}}, N_{a,t-1} \geqslant L | \mathcal{H}_{t-1}] \\ &\leqslant L + T d e^{-\frac{L r_a^2}{2\sigma^2}}.\end{aligned}$$

La deuxième inégalité utilise l'observation que $\mathbb{P}[a_t = a | \mathcal{H}_{t-1}]$ est fixe étant donné $\mathcal{H}_{t-1}$. La dernière inégalité utilise le lemme 12. En utilisant $L = 2\sigma^2 \frac{\ln(dT\Delta_a^2)}{r_a^2}$, nous obtenons

$$(C) \leqslant 2\sigma^2 \frac{\ln(dT\Delta_a^2)}{r_a^2} + \frac{1}{\Delta_a^2}.$$

### 4.5.2 Preuve du théorème 3

Par définition des rayons de préférence, pour une fonction de préférence linéaire (exemple 1), Chebyshev (exemple 2) ou $\varepsilon$-contrainte (exemple 3), nous pouvons prendre $\rho_\star = \rho_a = \frac{\Delta_a}{2}$ et $r_a = \frac{\Delta_a}{6}$. En utilisant ces valeurs dans la proposition 1, nous obtenons le théorème 3 :

$$\mathbb{E}[\mathfrak{R}(T)] \leqslant \sum_{\substack{a \in \mathcal{A} \\ a \neq \star}} \Big[(8C(d) + 24d + 18 + 72\sigma^2)(1+\sigma)^2 \frac{\ln(dT\Delta_a^2)}{\Delta_a} + \frac{4}{\Delta_a}\Big].$$

Soit $\Delta_a = \delta_a \sqrt{\frac{dN \ln N}{T}}$, pour $\delta_a \in (0, \sqrt{\frac{T}{dN \ln N}}]$. Le regret est borné par

$$\mathbb{E}[\mathfrak{R}(T)] \leqslant (8C(d) + 24d + 18 + 72\sigma^2)(1+\sigma)^2 \frac{\sqrt{NT}\ln(d^2 N \ln N)}{\delta_a \sqrt{d \ln N}} + \frac{4\sqrt{NT}}{\delta_a \sqrt{d \ln N}}$$

avec $\sigma^2 \leqslant 1/(4d)$, ce qui correspond à un ordre $\mathcal{O}(\sqrt{dNT}\ln d + \sqrt{dNT \ln N})$. Plus précisément, pour tout $d \leqslant \ln N$, la borne sur le regret est d'ordre $\mathcal{O}(\sqrt{dNT \ln N})$.

## 4.6 Évaluation empirique

Nous présentons maintenant des résultats expérimentaux visant à appuyer les résultats théoriques précédents ainsi qu'apporter des compléments pertinents relativement à l'utilisation de TS-MVN dans le contexte de bandits multi-objectifs discuté dans ce chapitre. Nous avons généré aléatoirement une configuration à 10 actions et $d = 2$ objectifs, telle que l'espace des objectifs $\mathcal{X} = [0,1]^2$. Nous considérons un problème où les observations proviennent d'une

Tableau 4.1 – Observations attendues avec leur valeur de préférence et leur écart de sous-optimalité pour les deux fonctions de préférence. L'observation attendue de l'action optimale est présentée en gras.

| $\boldsymbol{\mu}$ | $f(\boldsymbol{\mu})$ | | $\Delta$ | |
|---|---|---|---|---|
| | Linéaire | $\varepsilon$-contrainte | Linéaire | $\varepsilon$-contrainte |
| $(0.56, 0.46)$ | 0.50 | 0.46 | 0.17 | 0.26 |
| $(0.75, 0.26)$ | 0.46 | 0.26 | 0.21 | 0.46 |
| $(0.34, 0.79)$ | 0.61 | 0.00 | 0.06 | 0.72 |
| $(0.67, 0.50)$ | 0.56 | 0.50 | 0.11 | 0.22 |
| $(0.70, 0.42)$ | 0.54 | 0.42 | 0.13 | 0.29 |
| $(0.54, 0.72)$ | 0.65 | **0.72** | 0.02 | 0.00 |
| $(0.49, 0.62)$ | 0.57 | 0.00 | 0.10 | 0.72 |
| $(0.13, 0.84)$ | 0.56 | 0.00 | 0.11 | 0.72 |
| $(0.78, 0.60)$ | **0.67** | 0.60 | 0.00 | 0.12 |
| $(0.63, 0.44)$ | 0.51 | 0.44 | 0.16 | 0.28 |

distribution multi-Bernoulli, $\mathcal{B}_d(\boldsymbol{\mu}_a)$ pour chaque action $a \in \mathcal{A}$, ainsi qu'un problème où les observations proviennent d'une distribution normale multivariée, $\mathcal{N}_d(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$, de covariance

$$\boldsymbol{\Sigma}_a = \begin{bmatrix} 0.10 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}, \quad \forall a \in \mathcal{A}.$$

Un échantillon $\boldsymbol{y} \sim \mathcal{B}_d(\boldsymbol{\mu})$ d'une distribution multi-Bernoulli $d$-dimensionnelle de moyenne $\boldsymbol{\mu}$ est tel que $z_i \sim \mathcal{B}(\mu_i)$. Les expériences sont menées avec la fonction de préférence linéaire

$$f(\boldsymbol{x}) = 0.4x_1 + 0.6x_2, \quad \boldsymbol{x} \in \mathcal{X},$$

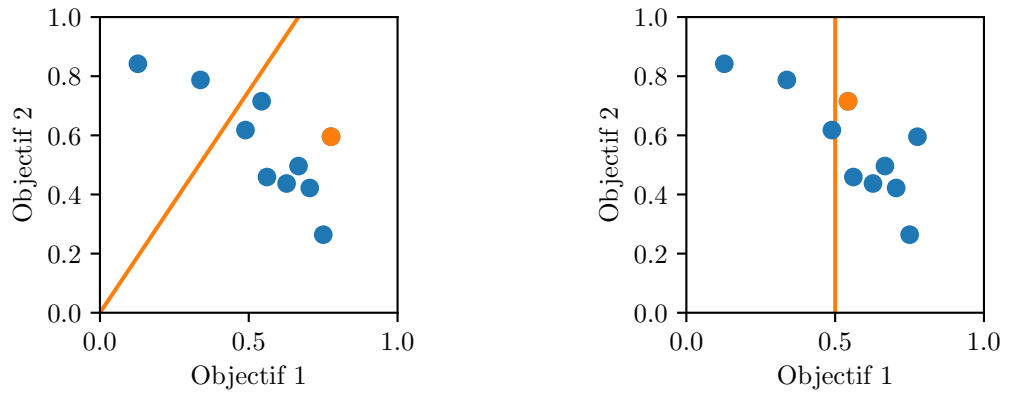
ainsi que la fonction de préférence $\varepsilon$-contrainte

$$f(\boldsymbol{x}) = \begin{cases} x_2 & \text{si} \quad x_1 \geqslant 0.5 \\ 0 & \text{sinon} \end{cases}, \quad \boldsymbol{x} \in \mathcal{X}.$$

Le tableau 4.1 montre les observations attendues pour toutes les actions avec leur valeur de préférence et leur écart de sous-optimalité associés pour les deux fonctions de préférence. La figure 4.5 montre les observations attendues et illustre les fonctions de préférence. Nous remarquons que l'action optimale n'est pas la même pour les deux fonctions de préférence. Toutes les expériences subséquentes sont réalisées sur 10 000 épisodes et répétées 100 fois. Les répétitions sont telles que le bruit $\boldsymbol{\xi}_t$ est le même pour toutes les approches comparées pour *la même répétition*. La performance des différentes approches est donc comparable pour une répétition donnée. Le but est de minimiser l'espérance du regret cumulatif (équation 4.1).

### 4.6.1 Validation de la théorie et pertinence du multi-objectif

Étant donné que la fonction de préférence est connue a priori, on pourrait être tenté d'aborder le problème sous la formulation des bandits traditionnels, mono-objectifs. Cela correspondrait




(a) Linéaire avec $\alpha_1 = 0.4$ et $\alpha_2 = 0.6$ (b) $\varepsilon$-contrainte avec $\ell = 2$ et $\varepsilon_1 = 0.5$

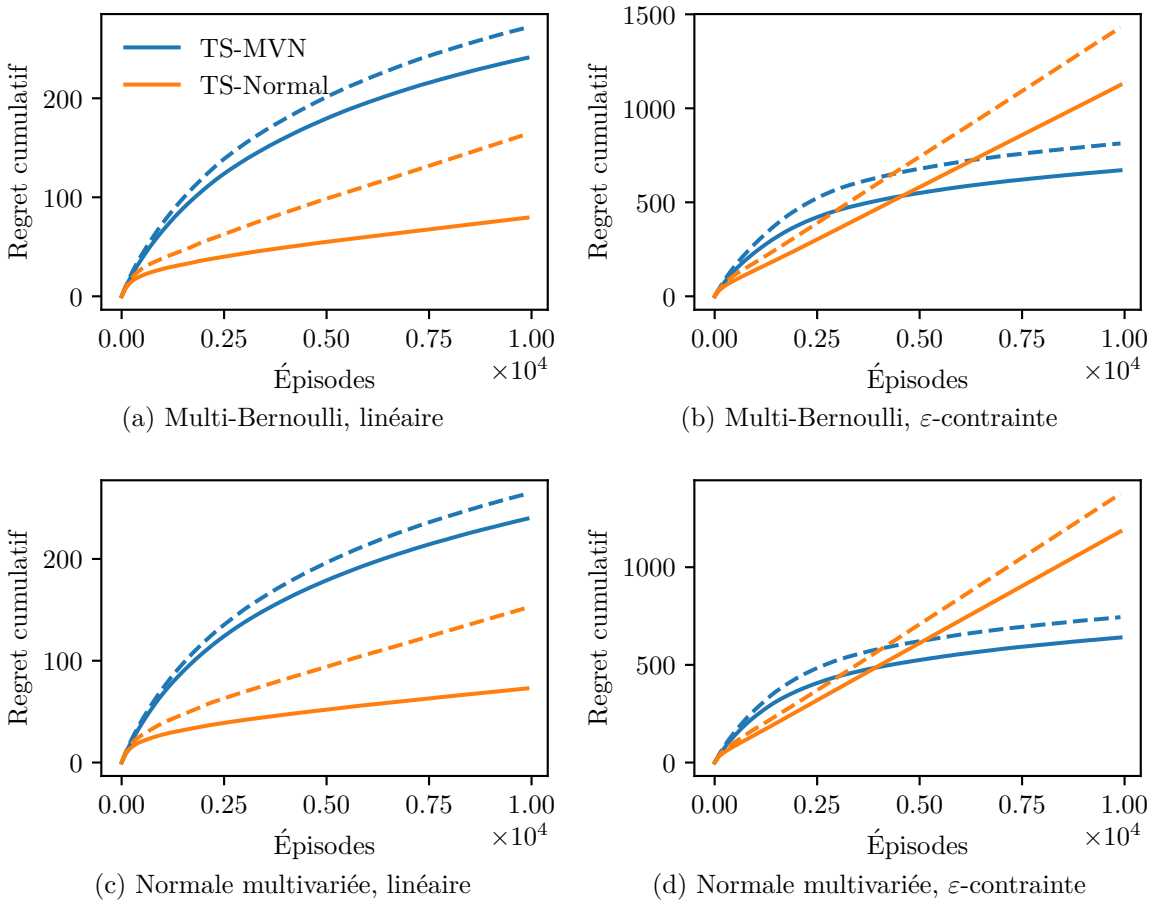
FIGURE 4.5 – Observations attendues pour les actions optimales (orange) et sous-optimales (bleu). À gauche, la ligne orange indique le gradient de la fonction de préférence linéaire et, à droite, la contrainte $\varepsilon_1$.

à optimiser par rapport à la valeur attendue de la fonction de préférence, $\mathbb{E}[f(\boldsymbol{y}_t)|a_t = a]$, au lieu d'optimiser $f(\boldsymbol{\mu}_a)$. Dans les expériences suivantes, nous comparons la performance du TS-MVN (algorithme 9) dans la formulation des bandits multi-objectifs (algorithme 8) avec TS d'a priori normal unidimensionnel (Agrawal and Goyal, 2013), TS-Normal, appliqué au problème des bandits multi-objectifs formalisé comme des bandits traditionnels (algorithme 2 où $y_t = f(\boldsymbol{y}_t)$).

La figure 4.6 montre le regret cumulatif de TS-MVN et de TS-Normal (dans la formulation de bandits traditionnels) pour les deux distributions d'observations et les deux fonctions de préférence. Nous remarquons que le taux de croissance sous-linéaire du regret cumulatif de TS-MVN concorde avec l'ordre des résultats théoriques présentés au théorème 3. Les résultats montrent également que, bien qu'il puisse être attrayant de ramener un problème multi-objectif à un problème mono-objectif par la fonction de préférence, ce n'est pas une bonne idée en pratique. Considérons la fonction de préférence $\varepsilon$-contrainte utilisée dans cette expérience. Elle est évaluée à 0 si $y_{t,1} < 0.5$, sinon à $y_{t,2}$. Avec des observations multi-Bernoulli, par exemple, cela signifie que $\mathbb{P}[f(\boldsymbol{y}_t) = 1] = \mu_{a_t,1}\mu_{a_t,2}$. Compte tenu de cela, $\arg\max_{a\in\mathcal{A}} f(\boldsymbol{\mu}_a) \neq \arg\max_{a\in\mathcal{A}} \mathbb{E}[f(\boldsymbol{y}_t)|a_t = a]$. Puisque l'action considérée comme optimale dans la formulation mono-objectif n'est pas la même que l'action optimale dans le problème multi-objectif, TS-Normal converge à la *mauvaise* action, ce qui explique le regret linéaire.

### 4.6.2 Utilisation de la covariance empirique

L'hypothèse de la connaissance de la matrice de covariance $\boldsymbol{\Sigma}_a$ dans l'algorithme 9 permet d'obtenir une garantie de convergence sur TS-MVN pour $\boldsymbol{\Sigma}_a = \boldsymbol{I}_d$. Cependant, cette configuration peut sembler sous-optimale puisqu'elle ne suppose aucune covariance entre les objectifs. Une alternative intuitive consiste à utiliser la matrice de covariance empirique, $\boldsymbol{\Sigma}_a = \widehat{\boldsymbol{\Sigma}}_{a,t-1}$,




(a) Multi-Bernoulli, linéaire

(b) Multi-Bernoulli, $\varepsilon$-contrainte

(c) Normale multivariée, linéaire

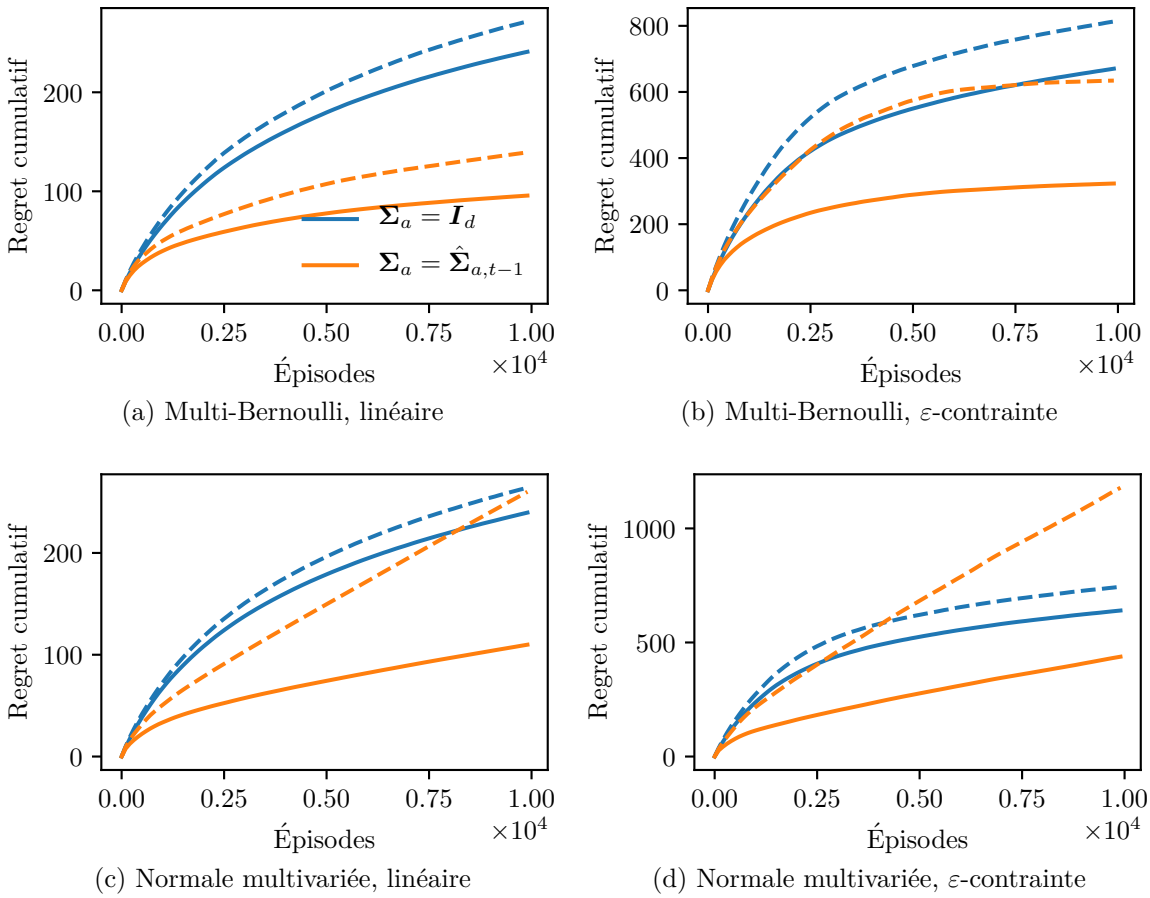
(d) Normale multivariée, $\varepsilon$-contrainte

FIGURE 4.6 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart type supérieur (ligne pointillée) avec TS-MVN et TS-Normal, pour les deux distributions d'observations et les deux fonctions de préférence.

avec

$$\widehat{\boldsymbol{\Sigma}}_{a,t} = \begin{cases} \frac{\sum_{s=1:a_s=a}^{t} \left(\boldsymbol{y}_s - \widehat{\boldsymbol{\mu}}_{a,t}\right)\left(\boldsymbol{y}_s - \widehat{\boldsymbol{\mu}}_{a,t}\right)^\top}{N_{a,t}-1} & \text{pour } N_{a,t} > 2 \\ \boldsymbol{I}_d & \text{sinon.} \end{cases}$$

La figure 4.7 montre le regret cumulatif de TS-MVN utilisant la covariance empirique comparativement à TS-MVN utilisant une covariance fixe correspondant à la matrice identité. On remarque que si l'utilisation de la covariance empirique peut sembler accélérer la convergence *en moyenne*, elles peuvent également mener à une divergence de TS-MVN (voir à la pire répétition sur les figures). En effet, l'utilisation de la matrice de covariance empirique *telle quelle* est risquée puisque fortement dépendante des observations initiales.




(a) Multi-Bernoulli, linéaire

(b) Multi-Bernoulli, $\varepsilon$-contrainte

(c) Normale multivariée, linéaire

(d) Normale multivariée, $\varepsilon$-contrainte

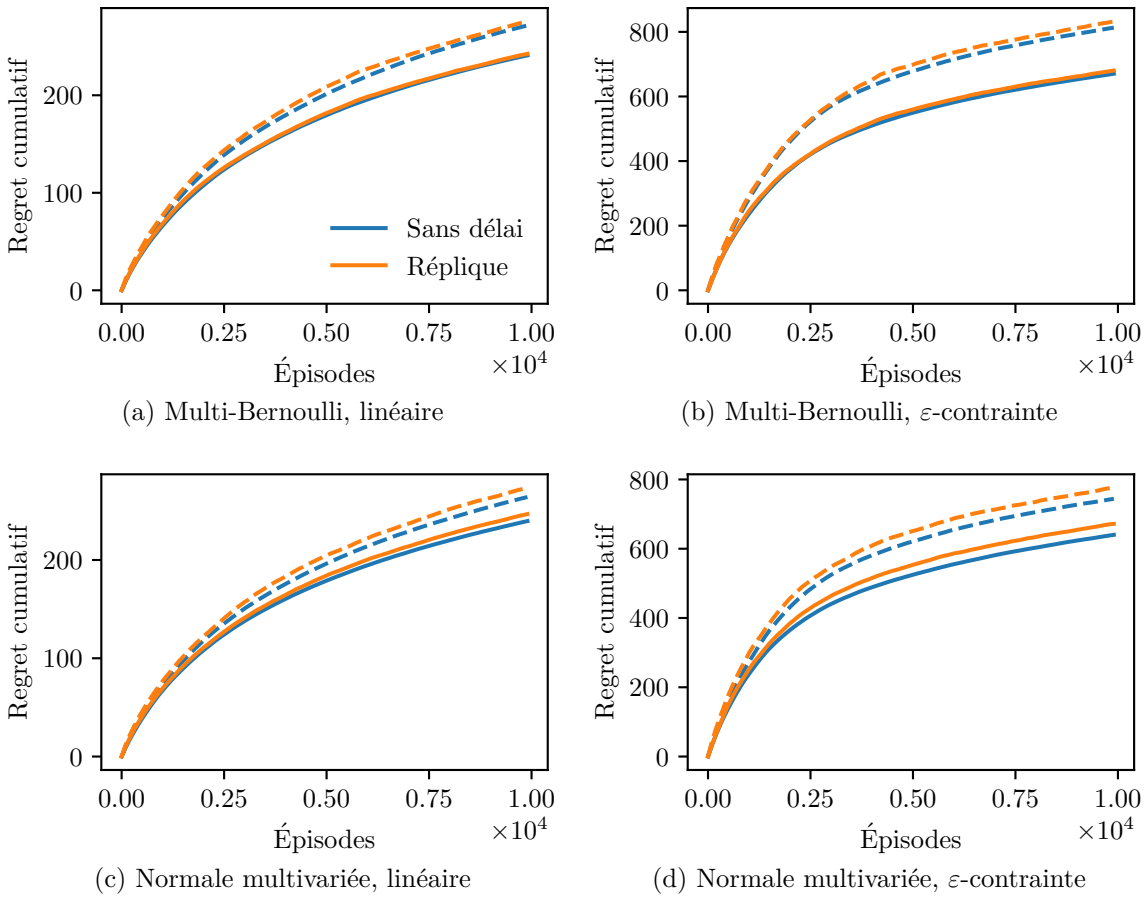
FIGURE 4.7 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart type supérieur (ligne pointillée) de TS-MVN avec $\boldsymbol{\Sigma}_a$ fixe et empirique, pour les deux distributions d'observations et les deux fonctions de préférence.

### 4.6.3 Rétroactions limitées de l'expert

Le fait qu'un expert soit intégré à la boucle d'apprentissage pour articuler la fonction de préférence peut être considéré comme une limitation puisque les actions d'un expert sont typiquement coûteuses (en temps). À cet effet, nous évaluons maintenant différentes approches pour faire face à la situation où l'expert ne fournirait des rétroactions qu'à tous les quelques épisodes. En d'autres termes, nous automatisons les rétroactions quand l'expert n'est pas disponible. À cet effet, nous considérons les techniques suivantes, où $\tau$ dénote l'épisode de la dernière rétroaction par l'expert. Dans les expériences qui suivent, l'expert offre des rétroactions à tous les 10 épisodes, ce qui correspond à $\tau = \max\{0, t - 10\}$. Soit l'action $a_\tau \in \mathcal{O}_\tau$.

**Réplique** Cette approche *préfère* les actions les plus récemment préférées par l'expert, soit $\mathcal{O}_t = \mathcal{O}_\tau$.

**Distance euclidienne carrée** Cette approche *préfère* l'action minimisant la distance euclidienne carrée à $a(\tau)$, soit $\mathcal{O}_t = \arg\min_{a \in \mathcal{A}} \|\boldsymbol{\theta}_{a,t} - \boldsymbol{\theta}_{a_\tau,\tau}\|_2^2$.




(a) Multi-Bernoulli, linéaire

(b) Multi-Bernoulli, $\varepsilon$-contrainte

(c) Normale multivariée, linéaire

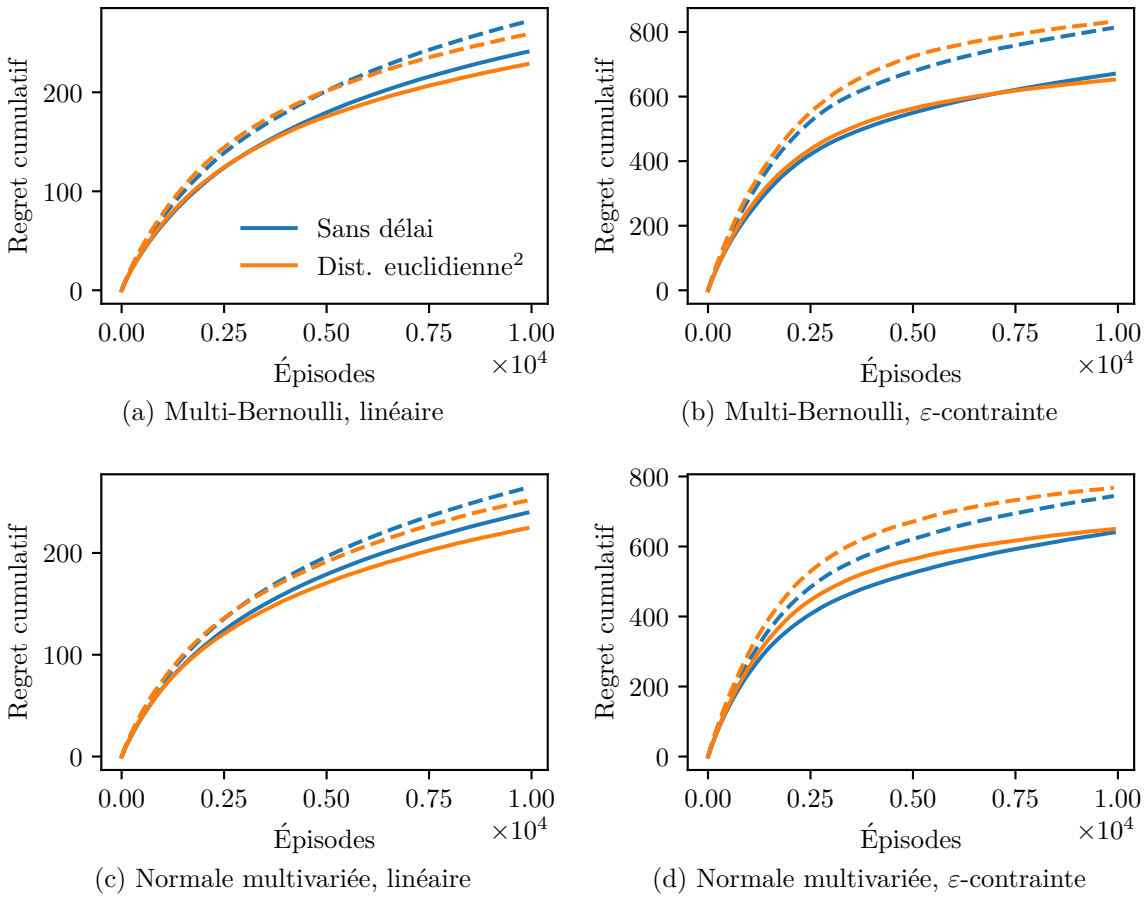
(d) Normale multivariée, $\varepsilon$-contrainte

FIGURE 4.8 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart type supérieur (ligne pointillée) de TS-MVN sans délai et avec délai géré par l'approche de réplique, pour les deux distributions d'observations et les deux fonctions de préférence.

**Hypothèse linéaire** Cette approche suppose que la fonction de préférence est linéaire. Ainsi, elle approxime la fonction de préférence par une fonction $\tilde{f}$ passant par $\boldsymbol{\theta}_{a_\tau,\tau}$, tel que $\mathcal{O}_t = \arg\max_{a\in\mathcal{A}} \tilde{f}(\boldsymbol{\theta}_{a,t})$.

Les figures 4.8, 4.9 et 4.10 comparent le regret obtenu avec TS-MVN avec et sans délai, pour les trois approches de gestion du délai mentionnées précédemment.

On remarque que l'approche de réplique (figure 4.8) ainsi que l'approche basée sur la minimisation de la distance euclidienne carrée (figure 4.9), quoique très simples, semblent en mesure d'atteindre une performance similaire au cas sans délai. Ces résultats sont très intéressants considérant que seulement 10% des rétroactions renvoyées à l'algorithme d'apprentissage proviennent de l'expert (les 90 autres % proviennent de l'approche de gestion du délai). Finalement, on remarque que l'approche basée sur l'hypothèse d'une fonction de préférence linéaire (figure 4.10) performe très bien lorsque la vraie fonction de préférence est également linéaire. En fait, la performance s'en trouve même améliorée comparativement au cas sans délai. Cela




(a) Multi-Bernoulli, linéaire

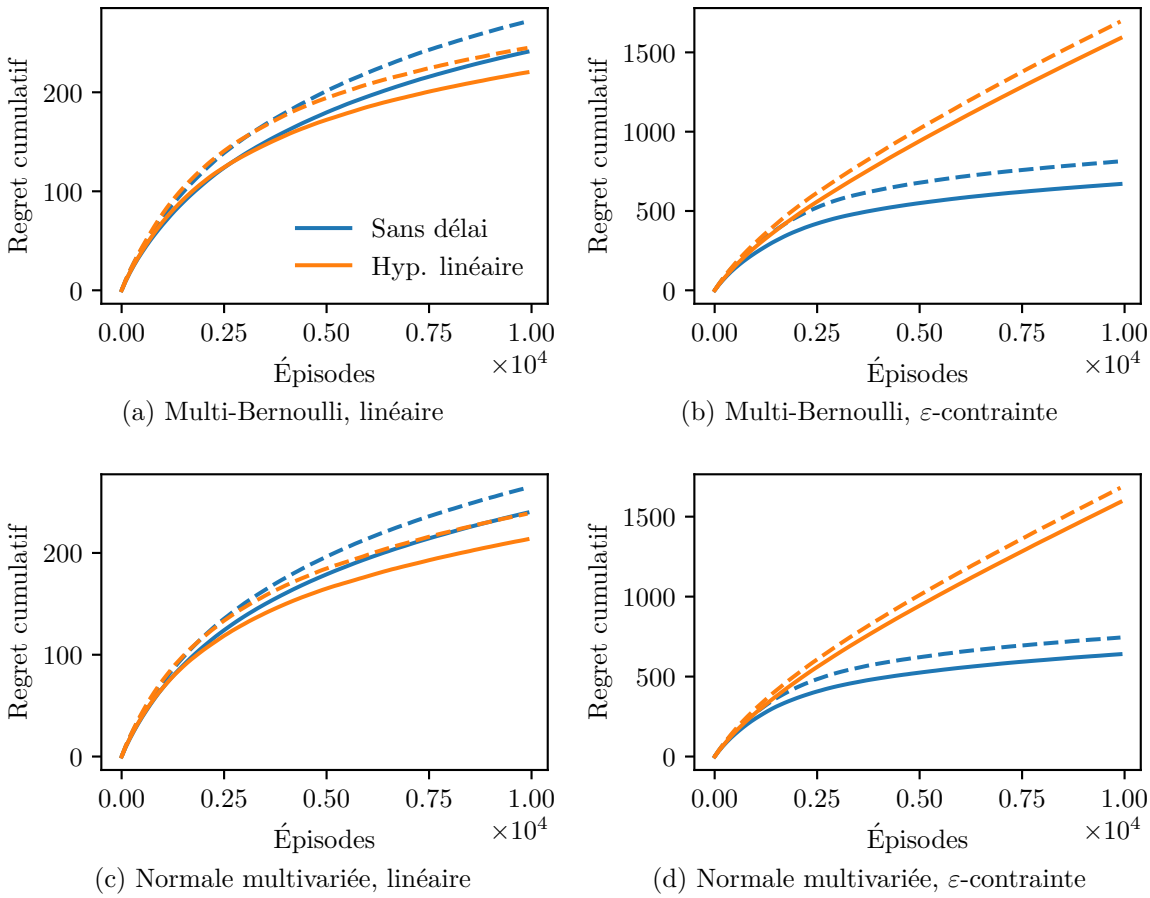
(b) Multi-Bernoulli, $\varepsilon$-contrainte

(c) Normale multivariée, linéaire

(d) Normale multivariée, $\varepsilon$-contrainte

FIGURE 4.9 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart type supérieur (ligne pointillée) de TS-MVN sans délai et avec délai géré par l'approche de minimisation de la distance euclidienne carrée, pour les deux distributions d'observations et les deux fonctions de préférence.

peut sembler surprenant et s'explique par le fait que la fonction linéaire hypothétique passe par la dernière option *préférée*. Considérons le cas où $a_\tau = \star$ et $\boldsymbol{\theta}_{a_\tau,\tau} = \boldsymbol{\mu}_a$. La figure 4.11a montre la fonction $\tilde{f}$ résultante passant par $\star$. L'écart de sous-optimalité calculé avec $\tilde{f}$ est de 0.08, comparativement à 0.02 avec la vraie fonction $f$, ce qui rend l'action optimale plus facile à distinguer, permettant ainsi de réduire l'accumulation de regret. Cependant, l'hypothèse d'une fonction de préférence linéaire peut se révéler désastreuse lorsque la vraie fonction de préférence n'est pas linéaire. Par exemple, dans le cas de la fonction de préférence $\varepsilon$-contrainte (figure 4.5b), la fonction $\tilde{f}$ passant par $\star$ n'est pas maximisée par $\star$, comme montré sur la figure 4.11b. Comme TS-MVN converge vers l'action maximisant $\tilde{f}$, il converge vers une action sous-optimale, ce qui explique le regret linéaire observé aux figures 4.10b et 4.10d.




(a) Multi-Bernoulli, linéaire

(b) Multi-Bernoulli, $\varepsilon$-contrainte

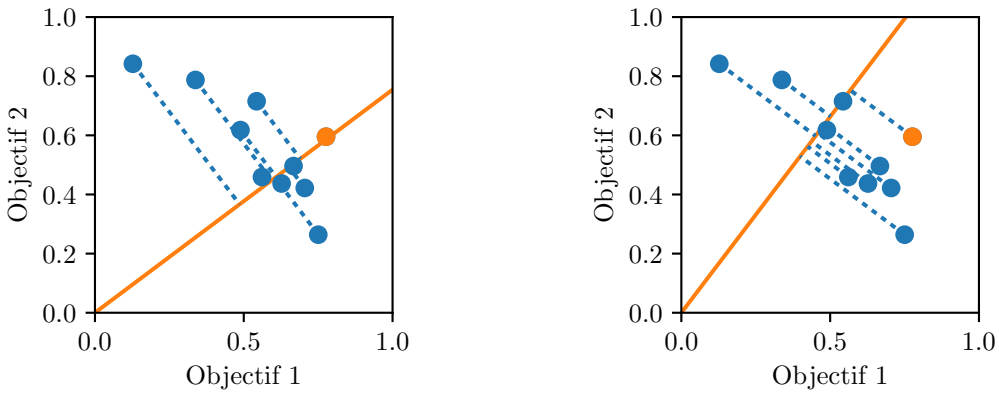
(c) Normale multivariée, linéaire

(d) Normale multivariée, $\varepsilon$-contrainte

FIGURE 4.10 – Pseudo-regret cumulatif moyen (ligne pleine) et avec un écart type supérieur (ligne pointillée) de TS-MVN sans délai et avec délai géré par l'hypothèse de fonction de préférence linéaire, pour les deux distributions d'observations et les deux fonctions de préférence.

## 4.7 Discussion

Dans ce chapitre, nous avons abordé le problème d'optimisation multi-objectifs en-ligne sous la forme des bandits multi-objectifs. Contrairement aux formulations précédentes, nous nous sommes concentrés sur le cadre a priori, où il existe une fonction de préférence à maximiser. Plus spécifiquement, nous avons abordé le problème dans lequel la fonction de préférence n'est pas connue (ou accessible). Nous supposons plutôt que nous disposons d'un utilisateur expert en mesure d'indiquer sa préférence parmi un ensemble d'options qui lui sont présentées. Nous avons introduit le concept de rayon de préférence pour caractériser la difficulté d'un problème multi-objectif via la robustesse de la fonction de préférence à la qualité des estimations disponibles. Nous avons montré comment cette mesure se rapporte à l'écart de sous-optimalité entre l'action optimale et l'action recommandée par un algorithme d'apprentissage. Nous avons utilisé ce nouveau concept pour fournir une analyse théorique de l'algorithme TS d'a priori normal multivarié dans le cadre multi-objectif (TS-MVN). Plus précisément, nous avons




(a) $\tilde{f}(\boldsymbol{x}) = 0.57x_1 + 0.43x_2$ (b) $\tilde{f}(\boldsymbol{x}) = 0.43x_1 + 0.57x_2$

FIGURE 4.11 – Actions optimales (orange) et sous-optimales (bleu) pour la fonction de préférence linéaire hypothétique illustrée (ligne orange).

fourni des garanties théoriques pour trois familles de fonctions de préférence.

En considérant une matrice de covariance prieure diagonale ($\boldsymbol{I}_d$), TS-MVN peut être vu comme l'utilisation d'une instance indépendante de TS d'a priori normal pour chaque objectif. Cela suggère que d'autres variantes du problème de bandits pourraient être abordées dans la situation multi-objectifs à l'aide de leur propre variantes de TS. Par exemple, le chapitre 5 présente une extension de Kernel TS pour les bandits structurés, introduit au chapitre précédant, à plusieurs objectifs.

Si l'introduction d'un humain expert dans la boucle d'apprentissage apparaît comme une avenue prometteuse pour articuler des préférences complexes (Christiano et al., 2017), cette approche présente certains défis avant d'être largement applicable. Par exemple, les interactions avec un expert peuvent se révéler coûteuses en temps et en ressources. De plus, la prise de décisions à répétition entraîne une dégradation des choix effectués. Ce phénomène connu sous le nom de *decision fatigue* (Vohs et al., 2008) et lié à l'*ego depletion* (Baumeister et al., 1998) a été remarqué et étudié dans différents contextes (Miller et al., 2010; Danziger et al., 2011). Il est donc impératif de réfléchir à l'impact de phénomène sur les décisions prises par un expert dans un système d'optimisation. À cet effet, la thèse survole différentes approches pour réduire le nombre de requêtes décisionnelles effectuées auprès de l'expert. Cela ouvre la porte à des travaux (Krueger et al., 2016) visant à explorer et caractériser les stratégies permettant de trouver un équilibre entre les requêtes à l'expert, leur coût ainsi que leur impact sur les décisions ultérieures.

Laisser reposer la sélection d'actions sur les choix d'un expert parmi des ensembles d'options amène également des défis en ce qui concerne la prise de décisions multi-critères (Köksalan et al., 2011). Il est connu que la comparaison de vecteurs multidimensionnels est une tâche fastidieuse et que des outils de visualisation graphiques peuvent supporter un expert dans la

prise de décision multi-critères (Miettinen, 2014). Si le problème peut sembler trivial pour une situation bi- ou tri-objectifs, il en est autrement quand le nombre d'objectifs augmente. Le support d'une dimensionnalité des objectifs plus importante repose donc à la fois sur le développement d'algorithmes plus robustes à la dimensionnalité ainsi que sur la mise en place d'outils permettant à un expert de prendre des décisions rapidement et efficacement, de manière en-ligne, en considérant plusieurs critères simultanément.

# Chapitre 5

# Optimisation de paramètres d'imagerie

Dans ce chapitre, nous abordons le problème d'optimisation de paramètres d'imagerie dans lequel le but est d'optimiser la paramétrisation d'un microscope à super-résolution de type STED, dans des conditions biologiques variables. L'approche traditionnelle consiste à effectuer une recherche en grille dans l'espace des paramètres de configuration possibles. Une telle couverture uniforme de l'espace entraîne la prise de plusieurs images inutilisables, en raison de leur *mauvaise* paramétrisation. Dans cette application, nous visons plutôt à trouver la configuration optimale tout en maximisant la qualité des images prises durant la recherche. Cela donne lieu à un compromis entre l'exploration de paramètres méconnus, à la recherche de paramètres potentiellement *meilleurs*, et l'exploitation de paramètres permettant d'obtenir des images d'une certaine qualité avec une certaine confiance. Il est donc naturel de formaliser cette application comme un problème de bandits. Plus spécifiquement, nous suggérons de tirer avantage de la structure sous-jacente aux actions (les paramètres appartenant à l'espace réel) et de formaliser ce problème dans le cadre des bandits structurés. L'optimisation peut alors être effectuée avec les approches introduites au chapitre 3.

Cependant, dans plusieurs situations, il n'est pas suffisant de se limiter à l'optimisation de la qualité des images. Différents objectifs, potentiellement contradictoires, doivent être pris en considération. Par exemple, lors de l'acquisition d'une image, la tentative d'augmenter la qualité d'une image peut entraîner une dégradation des structures et des marqueurs fluorescents, dont les effets sont observables dans cette même image. Il est alors nécessaire d'effectuer un compromis entre la qualité de l'image acquise et la préservation de l'échantillon. Nous montrons comment l'optimisation à plusieurs objectifs peut être dirigée en introduisant un utilisateur expert dans la boucle d'apprentissage, comme présenté au chapitre 4.

## 5.1 La microscopie STED

En permettant d'observer des structures à l'échelle nanoscopique, la microscopie STED a permis d'améliorer la compréhension des interactions moléculaires à la base d'une réponse immunitaire dans le cerveau (Fritzsche et al., 2017) et d'effectuer des avancées dans la recherche, par exemple sur le VIH (Hanne et al., 2016). Son principe repose sur la désactivation sélective de marqueurs fluorescents pour surpasser la limite de résolution imposée par la diffraction (Hell and Wichmann, 1994) [1]. Les fluorophores sont des composés chimiques ayant la propriété de ré-émettre de la lumière chaque fois qu'ils sont stimulés (excités) par une longueur d'onde spécifique. Ces derniers peuvent être ajoutés à des drogues ou à des anticorps, lesquels vont coller naturellement à la structure que l'on désire observer, par exemple, un neurone, ou une cellule vivante.

En microscopie confocale classique, un laser d'excitation balaie séquentiellement la région observée, récoltant au fur et à mesure les photons émis par les fluorophores excités. La résolution est alors limitée par le rayon minimal du laser. En microscopie STED (Hell and Wichmann, 1994; Willig et al., 2006), le laser d'excitation est suivi par un laser de déplétion dont la longueur d'onde peut désexciter les fluorophores par le biais de l'émission stimulée. Ayant une forme de beigne, ce dernier est chargé d'inhiber la quasi-totalité de la fluorescence, ne conservant que les molécules fluorescentes au centre du beigne. La figure 5.1 illustre les *point spread function*s (PSFs) des faisceaux lasers en question, obtenus par l'imagerie de billes d'or (100 nm). En saturant le processus d'émission stimulée, il est possible d'obtenir une PSF de fluorescence résultante plus petite que celle imposée par la limite de la diffraction. La figure 5.2 montre les PSFs de fluorescence obtenues par l'imagerie confocale et STED de microsphères fluorescentes de type TetraSpeck (100 nm), ainsi que la superposition des deux images. On remarque que l'image obtenue en STED a une résolution beaucoup plus fine que l'image confocale. On retrouve également la forme de beigne sur la superposition des deux images. Finalement, la figure 5.3 montre des images de neurones obtenues par microscopie confocale et STED. On remarque que l'imagerie STED permet d'observer des structures internes invisibles en microscopie confocale classique.

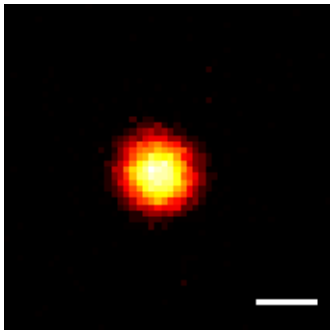
Le microscope STED peut être ajusté pour l'obtention d'images pertinentes pour les biologistes et les neuroscientifiques. Parmi les paramètres qui peuvent influencer les images acquises, on retrouve :

— la puissance du laser d'excitation ($\mu$W) ;

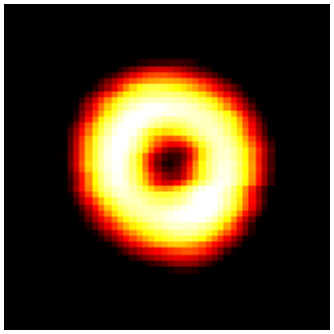
— la puissance du laser de déplétion (mW) ;

— le temps total passé à l'imagerie d'un pixel ($\mu$secondes), c'est-à-dire la somme des temps d'excitation et de déplétion.

1. Un prix Nobel a été remis à Stefan W. Hell en 2014 pour cette nouvelle technique de microscopie.

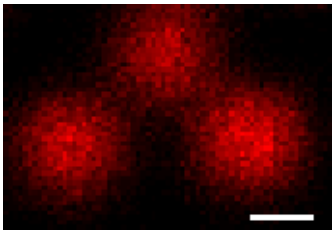


(a) Excitation (longueur d'onde de 488 nm)

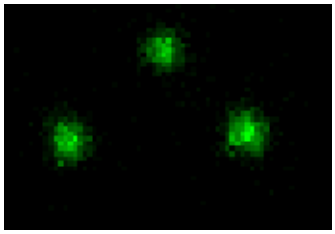


(b) Déplétion (longueur d'onde de 595 nm)

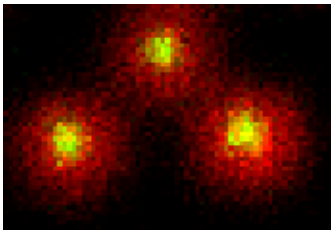
FIGURE 5.1 – PSFs des faisceaux lasers obtenues par l'imagerie de billes d'or. Le laser d'excitation provoque une émission de photons par les fluorophores alors que le laser de déplétion désactive cette émission. Échelle : 200 nm.



(a) Confocale



(b) STED

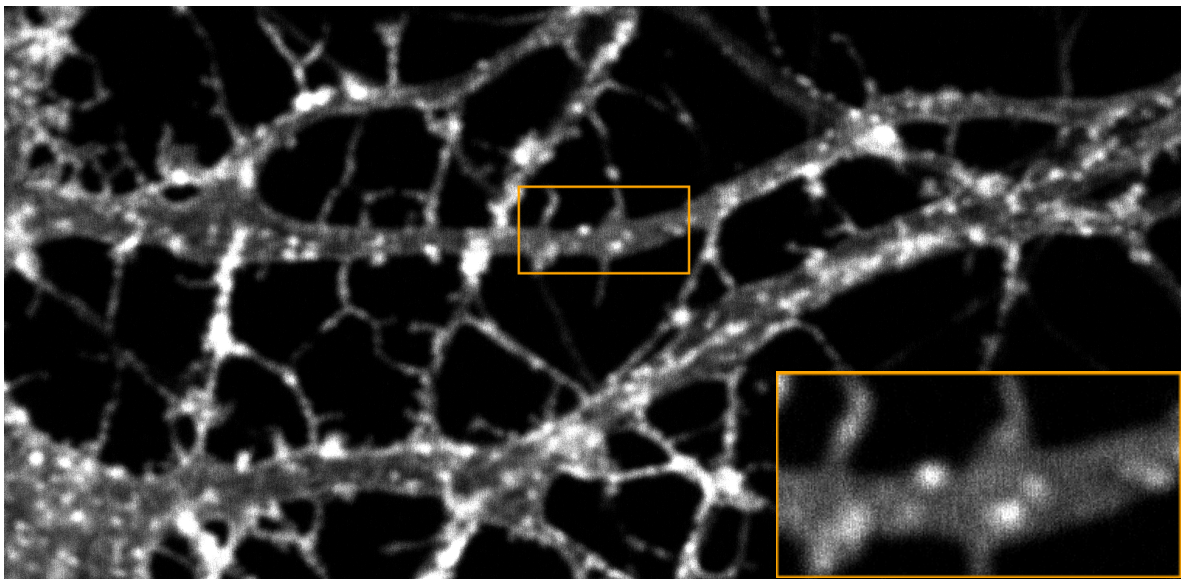


(c) Superposition

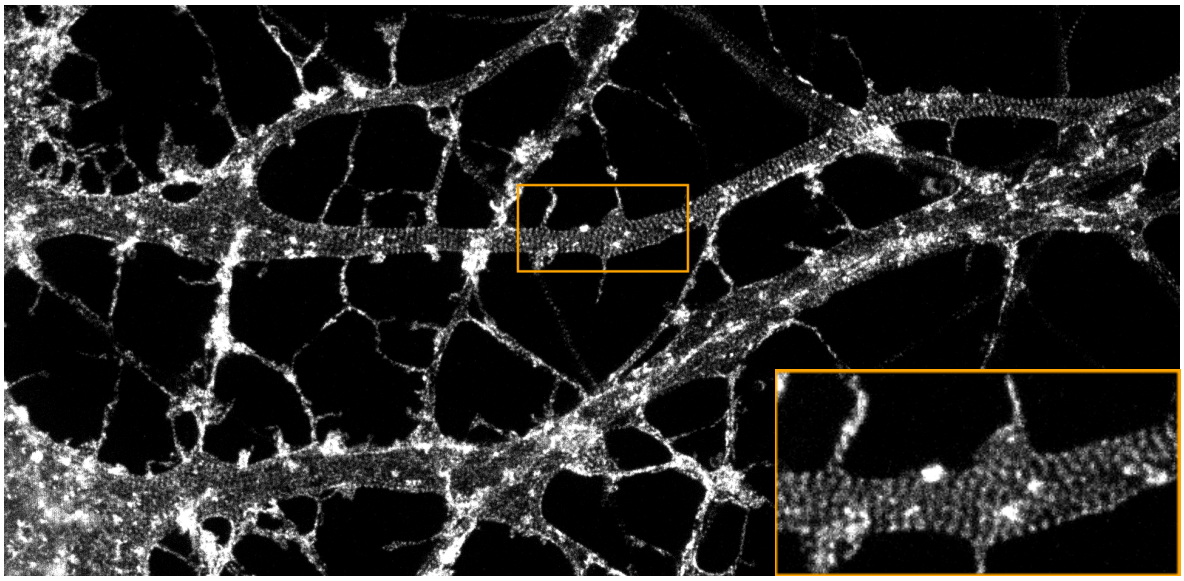
FIGURE 5.2 – PSFs de fluorescence de microsphères TetraSpeck. Échelle : 200 nm.

Les puissances d'excitation et de déplétion sont ajustées en modifiant le % de tension envoyée à un *accousto-optic modulator* (AOM), qui ajuste la puissance du laser. La figure 5.4 montre les puissances d'excitation et de déplétion en fonction du pourcentage de tension envoyée à l'AOM. L'ajustement de ces paramètres est un défi considérable puisque leur effet sur la *qualité* des images n'est pas linéaire et que leur impact est variable en fonction, par exemple, du type de fluorophore. De plus, les paramètres *optimaux* dépendent du type de structure observée ainsi que des différentes caractéristiques recherchées dans l'image.

La stratégie classique consiste à essayer chaque paramétrisation à plusieurs reprises, de manière à construire une estimation de la performance dans l'espace des paramètres. L'utilisateur responsable de l'imagerie choisit ensuite sa configuration *préférée* pour l'acquisition des images qui serviront réellement à l'analyse. L'approche proposée dans ce chapitre consiste plutôt à effectuer une optimisation en-ligne des paramètres de manière à maximiser l'obtention d'images utilisables pour l'analyse (répondant à certains critères à définir) simultanément à la recherche des paramètres. Contrairement à la stratégie classique, le but n'est pas d'identifier une configuration unique de paramètres à utiliser *pour toujours*, mais plutôt à poursuivre l'optimisation des paramètres simultanément à l'acquisition d'images pertinentes à l'analyse.

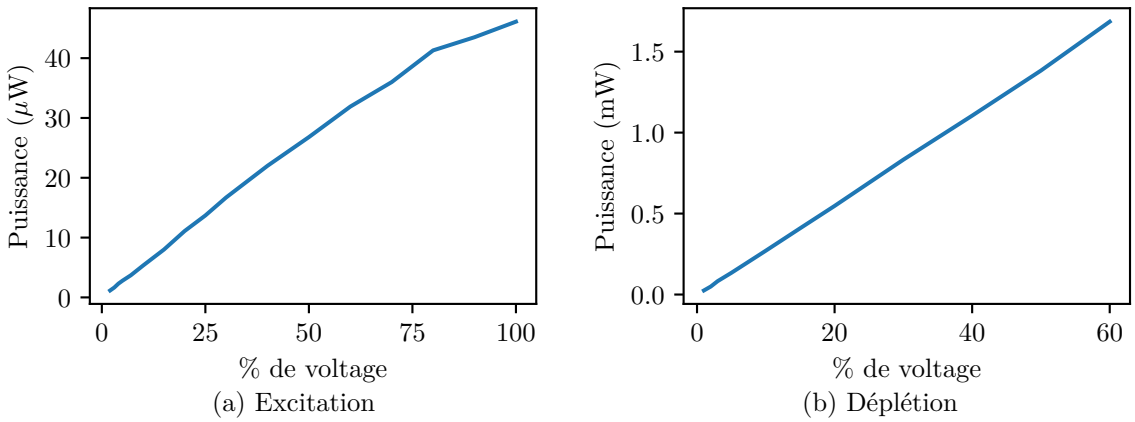


(a) Confocale



(b) STED

FIGURE 5.3 – Imagerie de la protéine d'actine sur des neurones hippocampaux fixés.



(a) Excitation (b) Déplétion

FIGURE 5.4 – Puissance des lasers en fonction du pourcentage de tension.

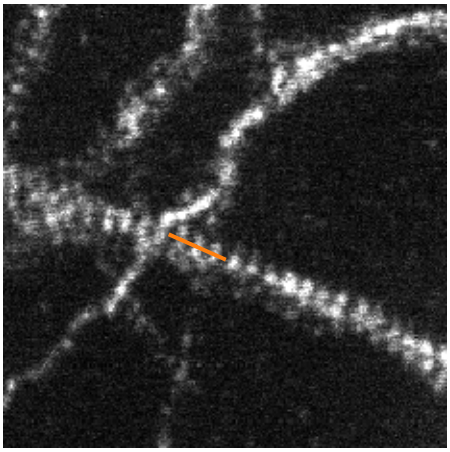
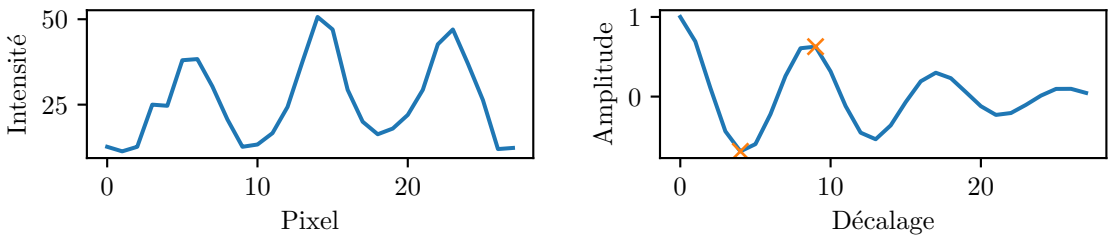


FIGURE 5.5 – Exemple d'anneaux d'actine (au long de la ligne orange) sur un axone.

La pertinence d'une image dépend du but motivant l'imagerie.

Par exemple, les expériences suivantes visent à optimiser en-ligne les paramètres d'un microscope STED pour l'observation d'anneaux d'actine sur des axones. L'axone est une projection du neurone servant au transport d'influx nerveux. Présente dans le cytosquelette, l'actine est une protéine essentielle, non seulement pour maintenir la structure des neurones, mais aussi pour la plasticité synaptique et l'interaction avec plusieurs autres protéines. Des fluorophores y sont attachés en utilisant la phalloïdine, soit une toxine issue de champignons. Les chercheurs souhaitent caractériser la présence d'anneaux d'actine dans les prolongements neuronaux, soient les axones et les dendrites. La figure 5.5 montre un exemple d'image STED sur laquelle apparaît une telle structure sur un axone.

Une approche typique pour caractériser la présence d'anneaux et leur clarté consiste à utiliser la mesure d'autocorrélation sur le signal le long d'une ligne donnée, typiquement tracée le long d'un prolongement qui *devrait* contenir des anneaux. La figure 5.6 montre un exemple de




(a) Intensité des pixels (b) Amplitude d'autocorrélation

FIGURE 5.6 – Exemple d'autocorrélation des anneaux dans un filament d'actine.

l'amplitude d'autocorrélation associée à la courbe d'intensité des pixels couverts par la ligne orange sur la figure 5.5. L'amplitude crête-à-crête maximale est utilisée comme métrique, correspondant à la différence d'amplitude entre les deux points marqués en orange dans l'exemple illustré par la figure 5.6b. Pour la suite, le terme « autocorrélation » désignera l'amplitude d'autocorrélation crête-à-crête maximale agin d'alléger le texte.

## 5.2 Optimisation des paramètres

Nous abordons le problème d'optimisation des paramètres sous l'angle des bandits structurés. Nous invitons les lecteurs à se référer au chapitre 3 pour une description détaillée du problème des bandits structurés ainsi que des approches qui seront discutées plus loin. Concrètement, nous cherchons à optimiser une fonction $f_\star : \mathcal{X} \mapsto \mathbb{R}$, où $\mathcal{X}$ correspond à l'espace des paramètres optimisés et $f_\star(x)$ correspond à la l'autocorrélation (moyenne) associée à la paramétrisation $x \in \mathcal{X}$. Une expérience se déroule sur un échantillon biologique spécifique. Cet échantillon est divisé en régions. Chaque épisode $t \in \mathbb{N}_{>0}$ correspond à l'acquisition d'une image, sur une région différente, avec la paramétrisation $x_t \in \mathcal{X}$. L'utilisateur est ensuite appelé à tracer trois lignes dans l'image, lesquelles servent à calculer l'autocorrélation $y_t$.

Dans une première expérience, nous visons à illustrer le potentiel de Kernel TS (algorithme 7, chapitre 3) dans à l'optimisation en-ligne des paramètres de STED comparativement à l'exploration uniforme classique. En raison de contraintes de temps, de main d'œuvre et de coût d'échantillons en laboratoires, l'exploration uniforme classique n'est jamais effectuée sur plus d'un paramètre à la fois. Ainsi, nous considérons dans cette première expérience un seul paramètre, soit la puissance du laser d'excitation. Pour des raisons de coût de calcul, Kernel TS doit travailler sur un sous-espace fini $\mathbb{X} \subset \mathcal{X}$. Cette contrainte n'est pas particulièrement problématique dans cette application puisque la procédure typique d'exploration uniforme requiert déjà un nombre fini et restreint de paramètres à comparer. Ainsi, nous considérons ici une discrétisation linéaire uniforme en 12 points des puissances d'excitation obtenues entre 3% et 38% de voltage, donc $|\mathbb{X}| = 12$.

Nous comparons l'exploration uniforme avec 5 échantillons par point (donc un total de 60

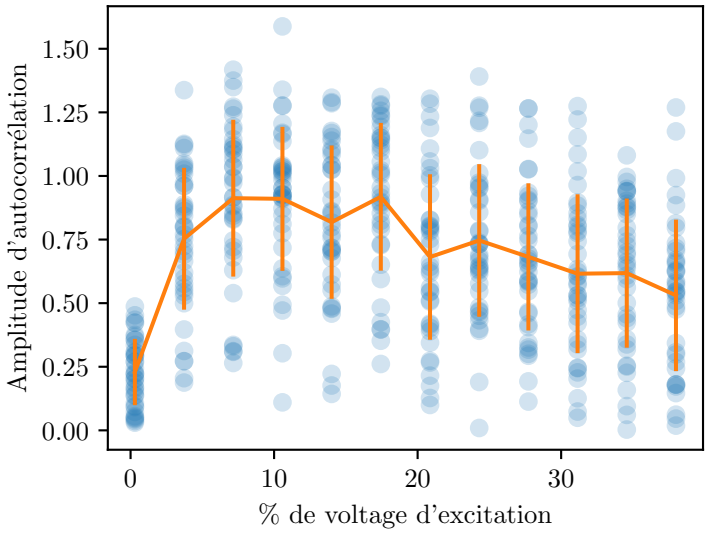



FIGURE 5.7 – Échantillons d'autocorrélation (bleu) obtenus en exploration uniforme de la puissance d'excitation, avec la moyenne et un écart type (orange).

échantillons) avec Kernel TS sur $T = 60$ épisodes. Suivant les résultats empiriques présentés précédemment (section 3.6.4), l'optimisation est effectuée avec Kernel TS sans inflation (algorithme 7, section 3.4), c'est-à-dire avec $v_t^2 = 1$. Un noyau gaussien (équation 1.26, chapitre 1) de bande-passante $\rho = 0.13$ est utilisé, ainsi que les bornes supérieure et inférieure suivantes sur le bruit, $\sigma_- = 1 \times 10^{-3}$ et $\sigma_+ = 0.3$. Ces paramètres ont été déterminés lors d'échange avec des experts disposant d'une connaissance a priori du domaine.

Pour débuter, 468 épisodes d'exploration uniforme initiale sur $\mathbb{X}$ ont permis d'obtenir 39 points par paramétrisation de $\mathbb{X}$. La figure 5.7 montre les données obtenues ainsi que l'espérance et l'écart type estimés à partir de ces points. Ces expériences ont permis à un expert d'identifier sur cette figure un ensemble de 4 paramètres optimaux, $\mathcal{O} = \{7.15, 10.58, 14.01, 17.44\}$ (% de voltage d'excitation). Dans le meilleur des cas, cet expert aurait souhaité n'imager qu'avec des paramètres parmi $\mathcal{O}$. Nous mesurons donc la performance correspondant à minimiser le regret cumulatif de *mauvais choix*, c'est-à-dire les choix de paramètres non-optimaux :

$$R^{\mathcal{O}}(T) \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}\{x_t \notin \mathcal{O}\}. \tag{5.1}$$

Étant donné l'importante variabilité qu'il peut y avoir entre deux échantillons biologiques, une technique par *bootstrap* est utilisée pour permettre une comparaison équitable entre les deux approches. Soit les observations $\mathcal{Y}_x = \{Y_{x,i}\}_{1 \leqslant i \leqslant 39}$ obtenues en évaluant l'autocorrélation au paramètre $x$. Si $Y_{x,i}$ est obtenue à l'épisode $t$, nous avons $Y_{x,i} = f_\star(x) + \xi_t$, où $\xi_t$ encode le bruit biologique, le bruit associé à la variabilité dans le choix des lignes pour l'évaluation de l'autocorrélation ainsi que le bruit associé à la région imagée. Soit $N_{x,t}$ le nombre de fois où le paramètre $x$ a été essayé jusqu'au temps $t$ (inclusivement). À l'épisode $t$, l'algorithme

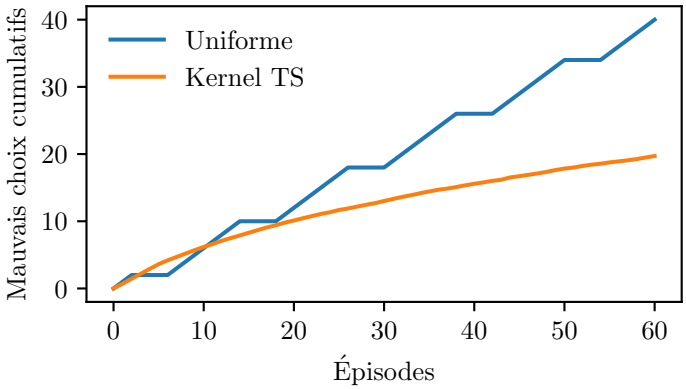



FIGURE 5.8 – Cumul de mauvais choix moyenné en optimisation de la puissance d'excitation.

choisit $x_t$ et observe $y_t = Y_{x_t, N_{x_t,t-1}}$. Les données $\mathcal{Y}_x$ sont ensuite mélangées aléatoirement 100 fois pour chaque $x \in \mathbb{X}$, permettant de générer 100 configurations différentes de séquences d'observations possibles. L'exploration uniforme ainsi que l'optimisation avec Kernel TS sont appliquées à chacune de ces 100 configurations. Ce modèle de simulation par bootstrap suppose évidemment que $N_{x,t} \leqslant 39$ pour tout $x \in \mathbb{X}, 1 \leqslant t \leqslant 39$, ce qui s'est révélé le cas dans les expérimentations effectuées.

La figure 5.8 montre le regret des mauvais choix (équation 5.1) obtenu avec la stratégie uniforme et Kernel TS. La courbe de regret associée à la stratégie uniforme présente une augmentation à tendance linéaire comme attendu, avec des plateaux causés par la sélection de paramètres optimaux, laquelle correspond à 1/3 des cas. On remarque que Kernel TS cumule un regret moyen sous-linéaire, tel qu'espéré.

En pratique, cependant, une exploration uniforme de l'ensemble des paramétrisations possibles ne sera pas effectuée conjointement à l'optimisation, le but étant de remplacer l'exploration uniforme par l'optimisation. Dans cette situation, nous suggérons de caractériser la performance d'un algorithme d'optimisation selon sa capacité à produire de *bonnes* images, c'est-à-dire des images pouvant être utilisées pas un biologiste ou neuroscientifique. Dans l'expérience considérée, l'acquisition d'images vise à observer des anneaux d'actine sur les axones (voir la figure 5.5). Nous distinguons alors les paramétrisations par leur capacité à produire des images sur lesquelles on peut apercevoir des structures en anneaux. Plus spécifiquement, nous utilisons une mesure de regret (à minimiser) basée sur l'accumulation de *mauvaises images*,

$$\mathfrak{R}^{\text{anneaux}}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}\{\text{anneaux non-visibles dans } i_t\}, \tag{5.2}$$

où $i_t$ correspond à l'image STED acquise avec la paramétrisation $x_t$. La figure 5.9 montre des exemples de mauvaises et bonnes images basées sur la présence d'anneaux visibles ou non.

La figure 5.10 montre le regret des mauvaises images (équation 5.2) obtenu avec la stratégie

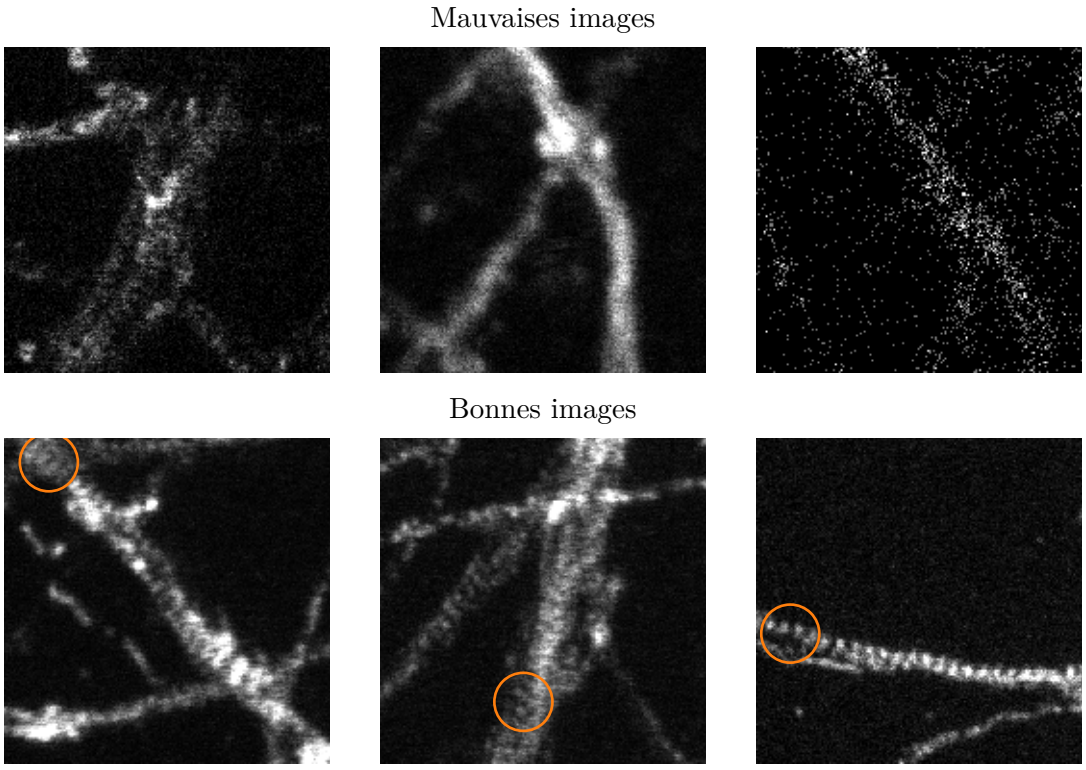



FIGURE 5.9 – Exemples de mauvaises et bonnes images (des anneaux sont encerclés).

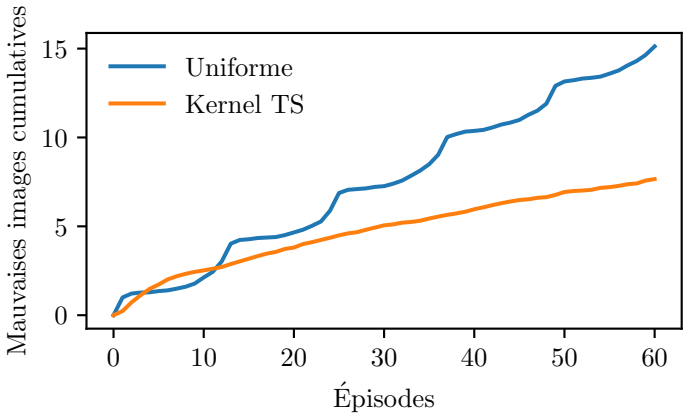



FIGURE 5.10 – Cumul de mauvaises images en optimisation de la puissance d'excitation.

uniforme et Kernel TS. La courbe de regret associée à la stratégie uniforme présente une augmentation à tendance linéaire tel qu'attendu, avec des variations et plateaux causés par la sélection de paramètres menant à de mauvaises images avec une certaine probabilité. Plus précisément, on observe que les plateau sur la figure 5.8 sont alignés avec des plateaux sur la figure 5.10. Contrairement au regret de mauvais choix (équation 5.1), le regret de mauvaises images (équation 5.2) contient les nuances associées à la probabilité de succès. On remarque que Kernel TS cumule toujours un regret moyen sous-linéaire, tel qu'espéré.

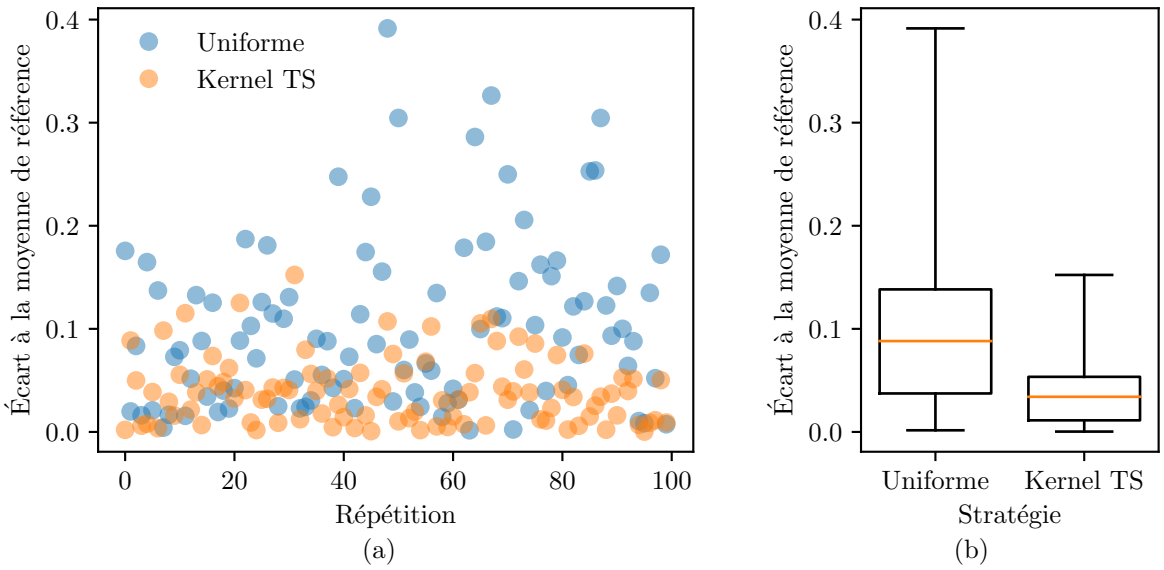



FIGURE 5.11 – Écart absolu entre l'estimation de l'action optimale et la moyenne de référence a) par répétition et b) par stratégie. Les boîtes couvrent du quartile inférieur au quartile supérieur, les poignées montrent l'étendue des données et la ligne orange indique la médiane.

Les experts en imagerie sont également intéressés par la capacité d'une stratégie à bien caractériser l'autocorrélation pour les paramètres optimaux. Comme Kernel TS tend à sélectionner de plus en plus de paramètres optimaux, il devrait être en mesure de mieux estimer le paramètre optimal sur lequel il converge qu'une stratégie uniforme qui dédierait seulement 5 essais à ce même paramètre. On remarque que c'est en effet le cas sur la figure 5.11, qui compare l'écart absolu à la moyenne de référence (voir la figure 5.7) pour le paramètre optimal le plus sélectionné avec Kernel TS avec l'écart obtenu en stratégie uniforme pour le même paramètre. Concrètement, soit $\tilde{\star}_t \overset{\text{def}}{=} \arg\max_{x \in \mathcal{O}} N_{x,t}$ la paramétrisation optimale sélectionnée le plus souvent par Kernel TS. Alors la figure 5.11 montre, pour chaque répétition, l'écart absolu entre la moyenne de référence et l'estimateur empirique de la moyenne

$$\widehat{\mu}_{\tilde{\star}_T,T} = \frac{1}{N_{\tilde{\star}_T,T}} \sum_{y \in \mathcal{Y}_{\tilde{\star}_T}} y$$

calculé sur les observations obtenues avec Kernel TS et de manière uniforme.

## 5.3 Optimisation à plusieurs objectifs

En pratique, maximiser l'autocorrélation seule n'est pas suffisant. Par définition, la mesure d'autocorrélation est influencée par la résolution obtenue sur l'image acquise (qui doit être assez grande pour bien distinguer les anneaux entre eux) ainsi que par le ratio signal à bruit (qui doit être assez grand pour bien distinguer les anneaux de l'arrière-plan). Malheureusement, en balayant l'échantillon pixel par pixel pour produire l'image STED, les lasers d'excitation

et de déplétion peuvent détruire la structure et les marqueurs fluorescents. Ces phénomènes, respectivement dénotés phototoxicité et photoblanchiment, tendent généralement à se produire davantage avec les paramètres d'imagerie maximisant à court terme la résolution et le ratio signal à bruit (donc l'autocorrélation). La phototoxicité et le photoblanchiment sont évidemment néfastes lors de l'acquisition d'une série d'images en une région fixe donnée, puisque les images se dégradent de plus en plus au long de la série. Cependant, leur effet se fait également sentir lors de l'acquisition d'une seule image : comme les lasers ont un diamètre non négligeable, l'imagerie d'un pixel entraîne une dégradation observable des pixels voisins. C'est la situation qui nous intéresse ici. L'imagerie de tissus fixés permet d'éviter la phototoxicité. Cependant, le problème du photoblanchiment demeure. L'expérience suivante vise donc à effectuer le compromis entre deux objectifs : la maximisation de l'autocorrélation et la minimisation du photoblanchiment.

### 5.3.1 Formulation du problème

Nous abordons ce problème en appliquant une logique similaire à celle proposée au chapitre 4 pour étendre les bandits structurés à la situation multi-objectifs. Soit $f_{\star,i} : \mathcal{X} \mapsto \mathbb{R}$ la fonction cible (inconnue) associée à l'objectif $i$. Nous utilisons la notation $\boldsymbol{f}_\star(x) = (f_{\star,1}(x), \dots, f_{\star,d}(x))$ pour dénoter l'évaluation de chaque fonction objectif au point $x$, tel que $\boldsymbol{f}_\star : \mathcal{X} \mapsto \mathbb{R}^d$. À chaque épisode $t \geqslant 1$, une action $x_t \in \mathcal{X}$ est appliquée et suivie d'une observation bruitée $\boldsymbol{y}_t = (y_{t,1}, \dots, y_{t,d})$. L'environnement est caractérisé par la fonction de préférence [2] $g : \mathbb{R}^d \mapsto \mathbb{R}$ d'un utilisateur expert. Idéalement, cet expert aimerait effectuer exclusivement des actions de l'ensemble optimal $\mathcal{O} \stackrel{\text{def}}{=} \arg\max_{x \in \mathcal{X}} g(\boldsymbol{f}_\star(x))$.

Un algorithme de bandits structurés multi-objectifs est une méthode (possiblement randomisée) pour sélectionner la prochaine action à entreprendre étant donné l'historique des choix antérieurs et des observations obtenues, $\mathcal{H}_t = \{x_s, \boldsymbol{y}_s\}_{s=1}^t$. Typiquement, il est supposé que l'algorithme propose une estimation $\boldsymbol{\theta}_{x,t}$ pour chaque action $x$ au temps $t$. Soit $\mathcal{O}_t \stackrel{\text{def}}{=} \arg\max_{x \in \mathcal{X}} g(\boldsymbol{\theta}_{x,t})$ l'ensemble des actions dont les estimations maximisent $g$. L'algorithme doit effectuer un compromis entre l'action $x_t \in \mathcal{O}_t$, qui maximise potentiellement $g$, et l'acquisition d'une observation supplémentaire pour un point relativement méconnu dans le but d'améliorer son estimation. L'algorithme 10 décrit ce problème de bandits structurés multi-objectifs.

**Remarque 19.** *L'algorithme 10 correspond à une extension directe de l'algorithme 8 (chapitre 4) à un espace d'actions structurées.*

2. Attention : la notation de cette fonction, dénotée $f$ au chapitre 4, est modifiée pour éviter la confusion avec la fonction d'espérance des observations sur l'espace des actions.

**Algorithme 10** Bandits structurés à multiples objectifs

À chaque épisode $t \geqslant 1$ :

1. L'agent observe $\mathcal{O}_t$.
2. L'agent sélectionne une action $x_t \in \mathcal{O}_t$.
3. L'agent observe $\boldsymbol{y}_t = \boldsymbol{f}_\star(x_t) + \boldsymbol{\xi}_t$, où $\boldsymbol{\xi}_t$ est un vecteur de variables aléatoires i.i.d.
4. L'agent met à jour ses estimations.

### 5.3.2 Kernel TS pour les bandits structurés multi-objectifs

Au chapitre 4, nous avons introduit l'algorithme Thompson *sampling* (TS) d'a priori normal multi-varié (MVN) pour les bandits multi-objectifs et avons présenté une analyse théorique pour une matrice de covariance diagonale entre les objectifs. Autrement dit, cela correspond à considérer une instance de TS d'a priori normal indépendantes pour chaque objectif. Similairement, nous étendons les bandits structurés à de multiples objectifs en traitant chaque objectif de manière indépendante. L'idée consiste à utiliser une instance de Kernel TS pour chaque objectif $i$, chargée de maintenir une distribution a posteriori sur $f_{\star,i}$ et servant à générer des options à présenter à un expert, *pour cet objectif*. Soit $\mathbb{X} \subset \mathcal{X}$ un sous-ensemble fini et discret de paramétrisations possibles. Soient les observations $\mathbf{y}_{t,i} = (y_{1,i}, \dots, y_{t,i})^\top \in \mathbb{R}^t$ obtenues pour l'objectif $i$ jusqu'au temps $t$ (inclusivement). À l'épisode $t$, pour chaque objectif $i$, la procédure consiste à :

1. estimer les bornes inférieures et supérieures sur le bruit, $\sigma_{-,t-1,i}$ et $\sigma_{+,t-1,i}$, avec $\mathbf{y}_{t,i}$ ;
2. calculer le paramètre de régularisation $\lambda_{t,i} \in \mathbb{R}_{>0}$ ;
3. calculer la moyenne postérieure ;

$$\widehat{\mathbf{f}}_{t-1,i} = \left(\mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda_{t+1,i}\boldsymbol{I}_t)^{-1}\mathbf{y}_{t-1,i}\right)_{x\in\mathbb{X}}; \tag{5.3}$$

4. calculer la covariance postérieure ;

$$\widehat{\boldsymbol{\Sigma}}_{t-1,i} = \frac{\sigma^2_{+,t-1,i}}{\lambda_{t,i}}\left(k(x,x) - \mathbf{k}_t(x)^\top (\mathbf{K}_t + \lambda_{t,i}\boldsymbol{I}_t)^{-1}\mathbf{k}_t(x)\right)_{x\in\mathbb{X}}; \tag{5.4}$$

5. échantillonner $\tilde{f}_{t,i} \sim \mathcal{N}(\widehat{\mathbf{f}}_{t-1,i}, \widehat{\boldsymbol{\Sigma}}_{t-1,i})$.

**Remarque 20.** *Les équations 5.3 et 5.4 correspondent respectivement à la moyenne et la variance postérieure de l'estimateur par régression à noyau à régularisation variable (théorème 1, chapitre 3).*

**Remarque 21.** *L'équation 5.4 utilise la borne empirique supérieure sur la variance du bruit plutôt que la connaissance du vrai bruit, comme Kernel TS standard (algorithme 7, chapitre 3.*

**Remarque 22.** *La version considérée de Kernel TS n'utilise pas le gonflement de la variance suggéré par la théorie (voir la section 3.4) puisque les résultats empiriques semblent en défaveur de cette pratique (voir la section 3.6.4).*

À l'épisode $t$, une option $\boldsymbol{\theta}_{x,t} = (\tilde{f}_{t,1}(x), \ldots, \tilde{f}_{t,d}(x))$ associée à chaque paramétrisation $x \in \mathbb{X}$ est présentée à l'expert. Celui-ci indique alors son choix $\mathcal{O}_t$ basé sur les options montrées et une paramétrisation $x_t \in \mathcal{O}_t$ est sélectionnée pour obtenir la prochaine image. Dans le cas présent, l'expert a une préférence unique, c'est-à-dire $|\mathcal{O}_t| = 1$.

### 5.3.3 Expériences

Comme précédemment, une expérience se déroule sur un échantillon biologique spécifique. Cet échantillon est divisé en régions. Chaque épisode $t \geqslant 1$ correspond à l'acquisition d'une image avec la paramétrisation $x_t \in \mathcal{X}$ suivie par l'obtention d'une observation bruitée $\boldsymbol{y}_t = (y_{t,1}, y_{t,2})$. À cet effet, l'utilisateur est appelé à tracer des lignes dans l'image (par exemple, voir la figure 5.5) pour calculer l'autocorrélation correspondant à $y_{1,t}$. Pour estimer le photoblanchiment, deux images confocales (d'impact photoblanchissant négligeable) sont prises, l'une précédant et l'autre suivant l'image STED. Soient $\bar{s}_t$ et $\bar{s}'_t$ le signal moyen sur l'avant-plan [3] des confocales précédant et suivant la STED à l'épisode $t$. Le photoblanchiment est estimé comme le ratio de perte de signal moyen et l'observation associée au second objectif est donnée par

$$y_{2,t} = \frac{\bar{s}_t - \bar{s}'_t}{\bar{s}_t}.$$

À la vue des résultats prometteurs obtenus précédemment, nous poursuivons avec l'optimisation simultanée de trois paramètres, soient les puissances d'excitation et de déplétion ainsi que le temps par pixel (voir la section 5.1). Nous considérons 10 puissances d'excitation linéairement sélectionnées dans l'intervalle de 0.5% à 40% de voltage, 10 puissances de déplétion linéairement sélectionnées dans l'intervalle de 5% à 60% de voltage et 10 valeurs de temps passé par pixel linéairement sélectionnées dans l'intervalle $[2 \times 10^{-6}, 60 \times 10^{-6}]$ (secondes). L'espace joint résultant, $\mathbb{X}$, comporte donc 1000 paramétrisations différentes.

Pour bien caractériser une combinaison de paramètres donnée en fonction de l'autocorrélation et du photoblanchiment, au moins deux mesures devraient être prises pour chaque paramétrisation de manière à compenser le bruit biologique et ainsi que le bruit dans les métriques d'objectifs. Étant donné le nombre de paramétrisations considérées (1000), une estimation à base de 5 observations par point impliquerait l'acquisition de 5000 images. Cela nécessiterait de 15 à 20 jours d'imagerie, répartis sur une période pouvant aller jusqu'à 6 semaines étant donné les contraintes expérimentales. Cela serait problématique étant donné la variation de la qualité des échantillons dans le temps (causée par une dégradation naturelle), laquelle implique une variation dans les paramètres optimaux. De plus, pour des échantillons vivants, cela impliquerait l'utilisation d'au moins 100 lamelles d'échantillons biologiques. Cela correspond au sacrifice de 5 à 10 cobayes animaux (bébés rats). Pour ces raisons, une exploration

3. L'avant-plan est trouvé par la méthoe d'Otsu (Otsu, 1979).

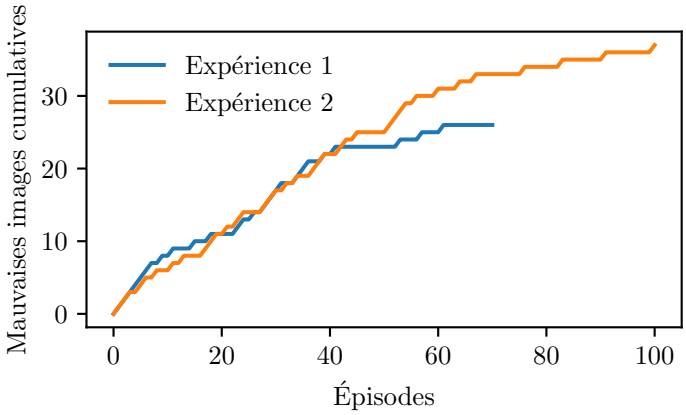



FIGURE 5.12 – Cumul de mauvaises images en optimisation multi-paramètres multi-objectifs.

uniforme serait considérée irréalisable dans des temps et des coûts respectables. Nous ne possédons donc pas de référence (*ground truth*) permettant de connaître l'ensemble de paramètres optimaux $\mathcal{O}$ ; il n'est donc pas possible de calculer le regret associé à la sélection des paramètres sous-optimaux (équation 5.1). L'extension proposée de Kernel TS pour des bandits structurés multi-objectifs est alors évaluée selon sa capacité à produire de *bonnes* images, c'est-à-dire que l'on vise à minimiser le regret des mauvaises images introduit précédemment (équation 5.2).

La figure 5.12 montre le regret cumulatif associé à l'acquisition de mauvaises images (équation 5.2) pour deux expériences (deux échantillons) d'horizon respectif $T = 70$ et $T = 100$. L'horizon était déterminé par le nombre de régions observables disponibles pour chaque échantillon. On remarque une convergence sous-linéaire tel qu'espéré suivant les garanties théoriques établies pour Kernel TS en bandits structurés (voir le chapitre 3) et TS-MVN en bandits multi-objectifs (voir le chapitre 4). Ces résultats préliminaires illustrent le potentiel des approches d'optimisation en-ligne pour l'ajustement automatique de paramètres d'imagerie STED pour l'acquisition d'images de bonne qualité.

## 5.4 Discussion

Dans ce chapitre, nous avons considéré le problème d'optimisation en-ligne des paramètres d'imagerie STED dans le cadre des bandits. Les paramètres faisant partie d'un ensemble caractérisé par une métrique de similarité, nous avons formulé le problème comme des bandits structurés et l'avons abordé avec les outils introduits précédemment au chapitre 3, plus spécifiquement l'algorithme Kernel TS. Les résultats expérimentaux montrent les bénéfices de l'optimisation en-ligne comparativement à la stratégie classique d'exploration uniforme. Ils appuient ainsi les intuitions de convergence de Kernel TS et illustrent son potentiel d'utilisation dans des applications réelles. Comme l'acquisition d'une image STED peut impliquer plusieurs objectifs (conflictuels) à optimiser, nous avons proposé une mécanique inspirée de

TS-MVN pour les bandits multi-objectifs, au chapitre 4. Nous avons ainsi étendu Kernel TS à la situation multi-objectifs. Les résultats préliminaires montrent que l'approche proposée converge vers des paramétrisations permettant d'acquérir des images présentant des structures pertinentes visibles.

### 5.4.1 Perspectives algorithmiques

L'application présentée dans ce chapitre est un bon exemple de cas réel qui ne s'inscrit directement dans aucune variante de bandits *telle quelle*. Ce type de problème doit plutôt être abordé sous le cadre joint de différentes variantes, ici les bandits structurés et multi-objectifs. Dans cette situation, les algorithmes propres aux différentes variantes doivent être adaptés pour satisfaire les nouvelles contraintes. Dans le cas présent, Kernel TS a été utilisé pour générer les options présentées à un expert humain dans une routine d'apprentissage à multiples objectifs. La perte des garanties de convergence est cependant un effet secondaire causé par de telles adaptations. Les résultats empiriques obtenus ici sont prometteurs et devraient motiver des analyses théoriques pour mieux comprendre l'approche résultante et encadrer sa performance par des garanties.

D'autres contraintes soulevées par l'application pourraient également être considérées. Par exemple, la possibilité de partager la connaissance acquise à travers différentes expériences pourrait permettre d'accélérer l'apprentissage. Cela serait très utile considérant que le nombre d'échantillons disponibles par expérience est assez limité, voire très limité sur des échantillons vivants. Ce problème pourrait être formulé sous le cadre des bandits contextuels (voir le chapitre 2), où l'image pourrait constituer le contexte. Les approches basées sur la régression à noyau n'étant pas adaptées à ce type de données, d'autres techniques, comme des réseaux de neurones (Allesiardo et al., 2014) ou des forêts aléatoires (Féraud et al., 2016), pourraient être considérées à cet effet.

L'application présentée dans ce chapitre est un exemple visant à illustrer le potentiel de l'optimisation multi-objectifs dans un contexte réel. Dans les faits, l'optimisation des paramètres d'imagerie à multiples objectifs pourra impliquer plus de deux objectifs. Par exemple l'utilisateur expert visant à acquérir des images de cellules vivantes (Berning et al., 2012) devra effectuer des compromis entre les mesures décrivant la qualité d'une image, comme l'autocorrélation, le ratio signal à bruit, la résolution de l'image et le photoblanchiment. L'imagerie de cellules vivantes impliquant typiquement des mouvements ou transformations que l'expert souhaite caractériser, des séries d'images devront être acquises de manière à *suivre* de tels changements. Cependant, les corps vivants réagissent à l'exposition lumineuse, un phénomène connu sous le nom de phototoxicité. Pour contrer cette réaction qui pourrait être confondue avec les changements naturels dans la structure, une approche consiste à réduire l'exposition aux lasers, ce qui est compensé par un temps d'imagerie plus long. Cependant, une vitesse d'imagerie trop lente ne permettra pas de bien observer les transformations ayant

lieu dans la structure. Il devient donc nécessaire de considérer le temps d'imagerie comme un objectif. De plus, plusieurs marqueurs fluorescents pourront également être utilisés simultanément et plusieurs images acquises au même moment (Winter et al., 2017). Les différents objectifs seront alors à optimiser conjointement sur plusieurs images. Cette situation impliquera rapidement un problème de présentation des données lié à l'intégration d'un expert humain pour la sélection d'une option *préférée*, tel que soulevé précédemment au chapitre 4. À terme, ce problème de visualisation pourrait constituer une limitation importante inhérente à la présence de multiples objectifs. Il pourrait donc devenir nécessaire de redéfinir de nouvelles métriques permettant d'encoder différents objectifs, par exemple, agglomérer l'autocorrélation, le ratio signal à bruit, la résolution et le photoblanchiment sous une métrique de *qualité* générale.

### 5.4.2 En route vers la microscopie intelligente

Les expériences présentées dans ce chapitre constituent un premier pas important dans l'automatisation des procédures d'imagerie en microscopie de pointe. Les algorithmes proposés permettent d'optimiser la configuration des systèmes STED dans des espaces trop complexes pour les procédures classiques. Par exemple, les résultats présentés en optimisation multi-objectifs (voir la section 5.3) n'auraient tout simplement pas pu être obtenus sans optimisation automatique. De plus, la nécessité d'effectuer un compromis entre différents objectifs conflictuels, sans la connaissance a priori de la fonction d'articulation des préférences, rendait ce problème difficilement abordable avec les méthodes existantes précédemment. Cette thèse a donc apporté des outils qui ont permis d'améliorer considérablement la procédure méthodologique en imagerie STED, plus précisément lorsque les experts sont confrontés à un échantillon biologique nouveau.

Ultimement, les approches d'apprentissage automatique pourraient permettre de reconsidérer la méthodologie typique stipulant que l'imagerie doit être effectuée de manière uniforme sur une région donnée. Des techniques ont déjà été proposées pour contrôler le photoblanchiment en éteignant les lasers après un certain temps étant donné le nombre de photons capturés (Staudt et al., 2011). Cependant, des méthodes intelligentes pourraient permettre d'ajuster le processus d'imagerie en fonction de la spatialité et de l'*intérêt* des structures. Cela soulève des questions relativement à la définition du concept d'intérêt. Des stratégies récentes (Balzarotti et al., 2016; Göttfert et al., 2017) visent à réduire le photoblanchiment et améliorer la résolution en effectuant du suivi de la structure. Elles ne permettent cependant pas encore de s'adapter à la biologie et nécessitent des infrastructures matérielles très spécifiques. Les approches basées sur l'optimisation ont le potentiel d'améliorer la performance des systèmes répandus dans les laboratoires de recherche.

Finalement, il est important de souligner que les approches proposées dans ce chapitre ne sont pas limitées à la microscopie STED. Plusieurs systèmes d'imagerie cellulaire ont le potentiel

d'être automatisés, notamment en ce qui a trait à l'analyse d'image (Hawkins, 2017; Kan, 2017).

# Conclusion

Ce dernier chapitre résume les contributions de la thèse et ouvre la voie à des travaux futurs. Rappelons que la thèse s'intéresse aux variantes du problème de bandits et à leurs applications pratiques à travers l'*adaptive clinical trial* (ACT) et l'optimisation de paramètres d'imagerie. L'annexe B présente d'autres travaux de recherche qui ont été réalisés durant le doctorat, mais qui ne sont pas couverts par cette thèse. L'annexe C présente la liste des publications ayant découlé de la thèse.

## Contributions

Les contributions résident à la fois dans des aspects algorithmiques, par l'extension d'approches de bandits traditionnels à différentes variantes de bandits, des aspects théoriques, par l'analyse de ces approches, ainsi que des aspects applicatifs, par un accent mis sur l'utilisation des approches proposées dans des applications à haut potentiel d'impact. Nous regroupons les contributions en quatre volets, correspondant respectivement aux chapitres 2 à 5.

### Les bandits contextuels

Dans un premier temps, nous nous sommes intéressés au problème des bandits contextuels, dans lequel les rétroactions sont obtenues avec un retard.

**Contribution 1** Nous avons proposé une extension de l'algorithme *best empirical sampled average* (BESA) aux bandits contextuels avec des actions disjointes via l'utilisation des *Gaussian processs* (GPs), GP BESA (algorithme 5). Contrairement aux approches de bandits contextuels existantes et basées sur l'algorithme *upper confidence bound* (UCB), GP BESA est randomisé plutôt que déterministe. Cela le rend plus attrayant dans la situation où le modèle est mis à jour avec un certain retard. En considérant la variance introduite par la régression sur un sous-ensemble des observations, GP BESA ouvre la porte à une utilisation différente des estimateurs par régression à noyau, ignorant la variance postérieure.

L'algorithme GP BESA a été utilisé pour l'allocation de traitements adaptative dans la phase d'ACT d'une expérience en laboratoire de collecte de données pour l'apprentissage de po-

litiques de traitement du cancer personnalisées. Un prolongement de la durée de vie allant jusqu'à 50% a été observé chez les cobayes animaux traités avec la stratégie de GP BESA comparativement aux stratégies classiques. Cela montre le potentiel des algorithmes de bandits pour l'ACT et devrait motiver davantage d'applications à des problèmes réels. La mise en pratique sur le terrain permet d'évaluer les algorithmes en dehors des hypothèses effectuées lors des analyses théoriques. Cela permet non seulement de définir de nouvelles contraintes et de nouveaux défis, mais encourage également le développement de méthodes d'apprentissage automatique robustes qui contribuent à l'avancement de domaines connexes.

### Les bandits structurés

Nous avons ensuite porté notre intérêt sur le problème des bandits structurés, dans lequel la fonction d'espérance de récompense, définie sur l'espace des actions, appartient au *reproducing kernel Hilbert space* (RKHS) d'un noyau donné, et dont les observations sont bruitées, d'un bruit de variance inconnue. Les méthodes existantes se basent sur l'hypothèse (très) forte d'une connaissance a priori de la variance du bruit, qui n'est généralement pas réalisable en pratique. Cette hypothèse est typiquement motivée par le fait que les intervalles de confiance des estimateurs par régression à noyau requièrent que le paramètre de régularisation soit constant. Comme la régularisation dépend naturellement du bruit, cela implique de travailler avec une borne supérieure constante sur le bruit (qui pourrait être très lâche) ou de supposer le bruit connu.

**Contribution 2** Pour nous attaquer à ce problème, nous avons d'abord proposé une extension des intervalles de confiance des estimateurs par régression à noyau, de manière à supporter une régularisation adaptative, variable dans le temps (théorème 1). Nous avons ensuite proposé une procédure d'ajustement automatiquement de la régularisation basée sur des estimateurs empiriques de la variance du bruit. Cela permet d'adapter la régularisation en fonction des observations obtenues au fil du temps, tout en maintenant des garanties théoriques sur l'intervalle de confiance de l'estimateur par régression (corollaire 1). La procédure résultante non triviale implique d'utiliser une estimation empirique de la borne inférieure sur le bruit pour calculer l'estimateur empirique de la borne supérieure. En permettant de lever la contrainte sur la connaissance a priori de la variance du bruit, ces outils rendent les approches basées sur la régression à noyau en-ligne beaucoup plus applicables dans la pratique.

**Contribution 3** Nous avons ensuite proposé une extension de l'algorithme Thompson *sampling* (TS) aux bandits structurés en RKHS, Kernel TS (algorithme 7). Nous avons effectué l'analyse théorique en utilisant les résultats de concentration proposés pour la régression à noyau de régularisation adaptative basée sur l'estimation empirique du bruit, ce qui nous a permis d'obtenir une borne sur le pseudo-regret cumulatif (théorème 2). Similairement aux analyses théoriques de TS linéaire (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016),

l'analyse de Kernel TS suggère d'échantillonner d'une distribution de variance gonflée par rapport à la variance postérieure. Nous avons étudié cette question par le biais d'expériences numériques. Les résultats ont appuyé nos intuitions, motivant des travaux futurs pour élucider la question du gonflement de la variance.

### Les bandits multi-objectifs

Nous avons également considéré le problème des bandits multi-objectifs, dans lequel il existe une fonction d'articulation des préférences qui n'est pas directement observable. Concrètement, nous supposons qu'un utilisateur expert est présent dans la boucle d'apprentissage et que ce dernier est en mesure d'indiquer sa préférence parmi un choix d'options qui lui sont présentées. Contrairement aux variantes précédentes, la formulation des bandits multi-objectifs tels qu'abordés ici est nouvelle et généralise d'autres variantes présentes dans la littérature.

**Contribution 4** L'introduction d'un expert (humain) dans la boucle d'apprentissage, sans observer directement la fonction de préférences, est un concept nouveau dans les bandits multi-objectifs. Nous avons donc introduit le concept des *rayons de préférence* (définition 6) pour mesurer la robustesse de la fonction d'articulation des préférences de l'expert face à différentes options qui lui sont proposées. Cet outil constitue un premier pas vers l'analyse et la compréhension de la prise de décision en-ligne à multiples objectifs.

**Contribution 5** Nous avons utilisé les rayons de préférence pour analyser l'algorithme TS d'a priori normal multivarié, TS-MVN (algorithme 9) de covariance diagonale. La technique de preuve étend l'analyse de TS d'a priori unidimensionnel (Agrawal and Goyal, 2013) au cas $d$-dimensionnel. Cela illustre bien la force du TS qu'est sa versatilité à travers ses a priori. Les rayons de préférences nous ont permis d'obtenir une borne sur le peusdo-regret cumulatif attendu (proposition 1), que nous avons spécialisée au cas de trois fonctions de préférence connues (théorème 3). L'analyse a permis de faire ressortir l'impact sur la difficulté du problème inhérente à l'estimation de plusieurs objectifs se traduisant par une tolérance au bruit inversement proportionnelle au nombre d'objectifs. Cette analyse pave également la voie à l'extension d'autres approches de bandits traditionnels vers des variantes multi-objectifs.

### L'optimisation de paramètres d'imagerie

Finalement, nous avons abordé le problème d'optimisation en-ligne des paramètres d'imagerie d'un système de microscopie STED. Comme les différentes configurations de paramètres appartiennent à un sous-espace des réels, nous avons formulé ce problème sous l'angle des bandits structurés et nous avons montré comment l'optimisation en-ligne avec Kernel TS permet l'acquisition d'images de qualité satisfaisante aux biologistes et neuroscientifiques, simultanément à la recherche d'une paramétrisation d'imagerie optimale.

En pratique, les utilisateurs experts sont généralement intéressés par un compromis entre différentes mesures caractérisant la *performance* de l'acquisition d'images. Nous avons donc formulé ce problème comme des bandits à la fois structurés et multi-objectifs.

**Contribution 6** Plus spécifiquement, nous avons proposé une extension de Kernel TS au multi-objectifs pour proposer des options à un utilisateur expert inclus dans la boucle d'apprentissage (algorithme 10). L'approche résultante a été déployée sur un vrai système STED et son utilisation a permis d'optimiser des paramètres dans des contextes d'utilisation où cela n'était pas possible avant. Cette application a permis de montrer comment divers outils introduits tout au long de la thèse ont pu être mis en commun pour la réalisation de cette application à impact non négligeable en imagerie super-résolution.

## Limitations et perspectives

En abordant différentes variantes des bandits multi-objectifs, cette thèse a contribué à l'avancement de la connaissance, à la fois par l'apport de nouveaux algorithmes que par ses analyses théoriques, le tout motivé par des applications concrètes. En considérant ainsi simultanément les aspects théoriques et les contraintes inhérentes aux applications, cette thèse vise à réduire l'écart entre la théorie et la pratique. Cela est effectué, par exemple, en levant les hypothèses de connaissance de la variance du bruit dans l'analyse des bandits structurés et de fonction de préférence accessible dans les bandits multi-objectifs. Les analyses présentées soulèvent cependant certaines questions demeurées non résolues dans la thèse. Le gonflement de la variance suggéré par l'analyse théorique de Kernel TS, similaire au gonflement suggéré par les analyses de TS linéaire (Agrawal and Goyal, 2014; Abeille and Lazaric, 2016), ne semble pas en accord avec les performances observées empiriquement. Cette question mériterait certainement d'être étudiée davantage. L'introduction d'un expert dans la boucle d'apprentissage en multi-objectifs a également soulevé plusieurs remarques, par exemple relativement à la difficulté de visualisation des données permettant à un humain d'effectuer un choix de préférence (Köksalan et al., 2011; Miettinen, 2014). Cette question n'a pas été abordée dans la thèse, mais constitue une limitation à considérer.

Le potentiel de GP BESA a été illustré par des simulations et son utilisation à l'ACT a permis d'optimiser la collecte de données sur des cobayes animaux. Similairement, les résultats obtenus en utilisant Kernel TS dans une boucle d'apprentisssage multi-objectifs basée sur les rétroactions d'un expert ont permis d'optimiser le processus d'imagerie. Cependant, autant GP BESA que l'extension multi-objectifs de Kernel TS ne possèdent présentement de garanties théoriques sur leur convergence. À moyen terme, il est nécessaire de fournir des analyses permettant d'encadrer la performance des algorithmes pour bien les comprendre, rendre leur amélioration possible et favoriser leur adoption (Chapelle and Li, 2011). Cela peut représenter un défi sous la contrainte d'hypothèses réalisables en pratique.

À travers les différents chapitres, le problème lié à l'extension de la dimensionnalité ainsi qu'à la quantité de données dans les approches basées sur la régression à noyau a été soulevé. Autant les algorithmes de bandits contextuels que structurés bénéficieraient d'extensions basées sur des approches de régression à meilleure gestion de la dimensionnalité (Djolonga et al., 2013; Wang et al., 2016; Li et al., 2016) ou à l'utilisation d'estimateurs alternatifs des distributions postérieures (Snoek et al., 2015; Springenberg et al., 2016). Ces nouvelles approches risquent de présenter des défis d'un point de vue théorique, mais seront nécessaires dans plusieurs situations où les données (ou les contextes) sont d'une nature qui les rend peu propices à être gérées par les techniques existantes. C'est le cas, par exemple, des images et des données biologiques telles les séquences d'ADN (Libbrecht and Noble, 2015).

D'un point de vue plus général, cette thèse illustre le potentiel applicatif des approches de bandits, tout en mettant l'accent sur le dilemme incessant entre la performance théorique et empirique. Elle soulève ainsi des discussions et encourage le perfectionnement des techniques sans oublier que l'on vise ultimement à résoudre des problèmes réels. Finalement, il faut garder en tête que le but des bandits consiste à mettre l'accent sur le compromis exploration-exploitation requis dans une dynamique d'apprentissage par renforcement en-ligne. Ainsi, les travaux futurs devraient également viser le transfert des connaissances développées dans les différentes variantes de bandits vers les problèmes, plus complexes, de prise de décisions séquentielle retrouvés en *reinforcement learning* (RL).

# Annexe A

# Quelques outils techniques

**Lemme 8** ((Anti-)Concentration d'une variable normale (Abramowitz and Stegun, 1964))**.** *Soit une variable $X$ échantillonnée d'une distribution normale de moyenne $\mu$ et de variance $\sigma^2$. Alors, pour tout $z \geqslant 1$,*

$$\frac{1}{2\sqrt{\pi}z} e^{-z^2/2} \leqslant \mathbb{P}[|X - \mu| > z\sigma] \leqslant \frac{1}{\sqrt{\pi}z} e^{-z^2/2}.$$

**Lemme 9** (Inégalité de Chernoff tirée de Rigollet (2015))**.** *Soient des variables $\sigma$-sous-gaussiennes i.i.d. $X_1, \ldots, X_N$ telles que $\mathbb{E}[X] = \mu$. Soit la moyenne empirique $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$. Alors, pour n'importe quel $\alpha \geqslant 0$,*

$$\mathbb{P}[|\widehat{\mu}_N - \mu| \geqslant \alpha] \leqslant 2e^{-\frac{N\alpha^2}{2\sigma^2}}.$$

**Lemme 10** (Inégalité de Chernoff-Hœffding)**.** *Soient des variables i.i.d. $X_1, \ldots, X_N$ contenues dans l'intervalle $[a, b]$ telles que $\mathbb{E}[X] = \mu$. Soit la moyenne empirique $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$. Alors, pour n'importe quel $\alpha \geqslant 0$,*

$$\mathbb{P}[|\widehat{\mu}_N - \mu| \geqslant \alpha] \leqslant 2e^{-\frac{2N\alpha^2}{(b-a)^2}}.$$

**Lemme 11** (Inégalité d'Azuma-Hœffding)**.** *Soit une super-martingale $(X_t)_{0 \leqslant t \leqslant T}$ correspondant à la filtration $\mathcal{H}_t$. S'il existe des constantes $c_t$ telles que $|X_t - X_{t-1}| < c_t$ pour tout $t = 1, \ldots, T$, alors pour n'importe quel $\alpha > 0$*

$$\mathbb{P}[X_T - X_0 \geqslant \alpha] \leqslant e^{-\frac{\alpha^2}{2\sum_{t=1}^T c_t^2}}.$$

**Lemme 12** (Inégalité de Chernoff $d$-dimensionnelle)**.** *Soient des variables $d$-dimensionnelles $\sigma$-sous-gaussiennes i.i.d. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ telles que $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ et $\widehat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i$. Alors, pour n'importe quel $a \geqslant 0$,*

$$\mathbb{P}[\widehat{\boldsymbol{\mu}}_N \succeq \boldsymbol{\mu} + a] = \mathbb{P}[\widehat{\mu}_{N,1} \geqslant \mu_1 + a) \wedge \cdots \wedge (\widehat{\mu}_{N,d} \geqslant \mu_d + a)] \leqslant e^{-\frac{dNa^2}{2\sigma^2}},$$

$$\mathbb{P}[\widehat{\boldsymbol{\mu}}_N \not\preceq \boldsymbol{\mu} + a] \leqslant \mathbb{P}[(\widehat{\mu}_{N,1} \geqslant \mu_1 + a) \vee \cdots \vee (\widehat{\mu}_{N,d} \geqslant \mu_d + a)] \leqslant de^{-\frac{Na^2}{2\sigma^2}},$$

$$\mathbb{P}[\widehat{\boldsymbol{\mu}}_N \notin B(\boldsymbol{\mu}, a)] \leqslant \mathbb{P}[(|\widehat{\mu}_{N,1} - \mu_1| \geqslant a) \vee \cdots \vee (|\widehat{\mu}_{N,d} - \mu_d| \geqslant a)] \leqslant 2de^{-\frac{Na^2}{2\sigma^2}}.$$

**Lemme 13** (Concentration d'une variable normale $d$-dimensionnelle)**.** *Soit une variable $X$ échantillonnée d'une distribution normale de moyenne $\mu$ et de variance $\sigma^2$. Du lemme 8 nous avons*

$$\mathbb{P}[|X-\mu|>z\sigma]\leqslant\frac{1}{2}e^{-z^2/2},$$

*Considérons maintenant le contexte multivarié dans lequel $\boldsymbol{X}$ dénote une variable aléatoire normale $d$-dimensionnelle de moyenne $\boldsymbol{\mu}$ et de covariance diagonale $\boldsymbol{\Sigma}$. Alors, pour tout $z\geqslant 1$,*

$$\mathbb{P}[\boldsymbol{X}\succ\boldsymbol{\mu}+z\sqrt{\operatorname{diag}(\boldsymbol{\Sigma})}]=\mathbb{P}[(X_1>\mu_1+z\sigma_1)\wedge\cdots\wedge(X_d>\mu_d+z\sigma_d)]\leqslant\left(\frac{1}{4}e^{-z^2/2}\right)^d,$$
$$\mathbb{P}[\boldsymbol{X}\not\prec\boldsymbol{\mu}+z\sqrt{\operatorname{diag}(\boldsymbol{\Sigma})}]\leqslant\mathbb{P}[(X_1\geqslant\mu_1+z\sigma_1)\vee\cdots\vee(X_d\geqslant\mu_d+z\sigma_d)]\leqslant\frac{d}{4}e^{-z^2/2},$$
$$\mathbb{P}[\boldsymbol{X}\notin B(\boldsymbol{\mu},z\sqrt{\operatorname{diag}(\boldsymbol{\Sigma})})]\leqslant\mathbb{P}[(|X_1-\mu_1|\geqslant z\sigma_1)\vee\cdots\vee(|X_d-\mu_d|\geqslant z\sigma_d)]\leqslant\frac{d}{2}e^{-z^2/2}.$$

**Lemme 14** (Anti-concentration d'une variable normale $d$-dimensionnelle)**.** *Soit une variable aléatoire normale $d$-dimensionnelle $\boldsymbol{X}$, de moyenne $\boldsymbol{\mu}$ et de covariance diagonale $\boldsymbol{\Sigma}$. Agrawal and Goyal (2013) proposent le résultat d'anti-concentration à partir d'Abramowitz and Stegun (1964) pour $d=1$ :*

$$\mathbb{P}[X>\mu+z\sigma]\geqslant\frac{z}{\sqrt{2\pi}(z^2+1)}e^{-z^2/2}.$$

*Étendu à $d\in\mathbb{N}$, nous avons*

$$\mathbb{P}[\boldsymbol{X}\succ\boldsymbol{\mu}+z\sqrt{\operatorname{diag}(\boldsymbol{\Sigma})}]=\mathbb{P}[(X_1>\mu_1+z\sigma_1)\wedge\cdots\wedge(X_d>\mu_d+z\sigma_d)]\geqslant\left(\frac{z}{\sqrt{2\pi}(z^2+1)}e^{-z^2/2}\right)^d.$$

**Lemme 15** (Formule de Sherman-Morrison-Woodbury)**.** *Soit la matrice inversible $\mathbf{A}\in\mathbb{R}^{n\times n}$ et les vecteurs colonnes $\mathbf{u},\mathbf{v}\in\mathbb{R}^{n\times k}$. En assumant que $(\boldsymbol{I}_k+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u})$ soit inversible, alors*

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^\top)^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{u}(\boldsymbol{I}_k+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u})^{-1}\mathbf{v}^\top\mathbf{A}^{-1}.$$

# Annexe B

# Travaux additionnels

Cette annexe présente d'autres travaux de recherche qui ont été effectués durant le doctorat, mais qui ne figurent pas dans les axes principaux couverts par cette thèse.

**Auto-complétion de requêtes à l'aide de bandits pour l'agrégation de moteurs recherche** L'aide aux utilisateurs en proposant des requêtes complètes au fur et à mesure de leur saisie est une caractéristique commune des systèmes de recherche connus sous le nom d'*auto-complétion des requêtes*. Un moteur d'auto-complétion automatique de requête peut utiliser différentes informations disponibles (par exemple, l'utilisateur est anonyme, l'utilisateur a un historique, l'utilisateur a visité le site avant la recherche ou non, etc.) afin d'améliorer ses recommandations. Il existe de nombreuses stratégies possibles pour l'auto-complétion de la requête et un défi consiste à concevoir un moteur optimal qui considère et utilise toutes les informations disponibles. Une stratégie alternative consiste à agréger plusieurs moteurs afin d'améliorer la diversité des recommandations en combinant la capacité de chaque moteur à digérer l'information disponible différemment, tout en conservant la simplicité de chaque moteur. L'objectif principal de cette recherche est donc de trouver un tel mélange de moteurs d'auto-complétoin des requêtes qui surpasserait un moteur unique. Nous abordons ce problème dans le cadre des bandits et évaluons quatre stratégies pour surmonter ce défi. Les expériences effectuées sur trois ensembles de données réels montrent qu'un mélange de moteurs peut en effet surpasser un moteur unique.

— Durand, A., Beaumont, J.-A., Gagné, C., Lemay, M., and Paquet, S. (2017a). Query completion using bandits for engines aggregation. In *3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

**Amélioration des bornes sur le regret de Pareto UCB1 en bandits multi-objectifs** Dans ce travail, nous introduisons une approche simple pour transposer facilement l'analyse de regret des heuristiques de bandits traditionnels vers la situation à multiples objectifs introduite par Drugan and Nowe (2013). En utilisant notre approche, nous améliorons l'al-

gorithme Pareto UCB1, c'est-à-dire l'extension multi-objectifs de l'UCB1, en effectuant une analyse de regret plus serrée. Le Pareto UCB1$^\star$ qui en résulte a également l'avantage d'être empiriquement utilisable sans aucune approximation.

— Durand, A., Bordet, C., and Gagné, C. (2014). Improving the Pareto UCB1 algorithm on the multi-objective multi-armed bandit. In *Workshop on Bayesian Optimization at the 27th Neural Information Processing Systems (NIPS)*.

**TS pour les bandits combinatoires et son application à la sélection de *features* en-ligne** Dans ce travail, nous abordons le problème des bandits combinatoire. Sous certaines hypothèses, nous suggérons de ramener le problème à des bandits traditionnels et d'utiliser l'algorithme TS. Nous appliquons cette stratégie au problème de sélection de *features* en-ligne et obtenons des résultats qui montrent le potentiel de TS. Nous discutons également les défis associés à la sélection de *features* en-ligne et suggérons des orientations pertinentes pour des travaux futurs.

— Durand, A. and Gagné, C. (2014). Thompson sampling for combinatorial bandits and its application to online feature selection. In *Proc. of the 28th AAAI Conference, Workshop on Sequential Decision-Making with Big Data*, pages 6–9.

# Annexe C

# Liste des publications

## C.1 Conférences et *workshops*

Durand, A., Beaumont, J.-A., Gagné, C., Lemay, M., and Paquet, S. (2017a). Query completion using bandits for engines aggregation. In *3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A. and Gagné, C. (2017b). Thompson sampling for user-guided multi-objective bandits optimization. In *3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A. and Pineau, J. (2015b). Cancer treatment optimization using gaussian processes. In *2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A. and Pineau, J. (2015a). Adaptive treatment allocation using sub-sampled gaussian processes. In *AAAI 2015 Fall Symposium on Embedded Machine Learning (EML)*.

Durand, A. and Pineau, J. (2015c). Treatment allocation as contextual bandit. In *Workshop on Machine Learning in Healthcare at the 28th Neural Information Processing Systems (NIPS)*.

Durand, A., Bordet, C., and Gagné, C. (2014). Improving the Pareto UCB1 algorithm on the multi-objective multi-armed bandit. In *Workshop on Bayesian Optimization at the 27th Neural Information Processing Systems (NIPS)*.

Durand, A. and Gagné, C. (2014). Thompson sampling for combinatorial bandits and its application to online feature selection. In *Proc. of the 28th AAAI Conference, Workshop on Sequential Decision-Making with Big Data*, pages 6–9.

## C.2 Journaux en préparation

Durand, A. and Gagné, C. (2017a). Estimating quality in user-guided multi-objective bandits optimization. *arXiv preprint arXiv:1701.01095*.

Durand, A., Maillard, O.-A., and Pineau, J. (2017b). Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv:1708.00768*.

# Bibliographie


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2312–2320.

Abeille, M. and Lazaric, A. (2016). Linear Thompson sampling revisited. *arXiv preprint arXiv :1611.06534*.

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. (2011). Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043.

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6) :1926–1951.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 39.1–39.26.

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 99–107.

Agrawal, S. and Goyal, N. (2014). Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv :1209.3352*.

Allesiardo, R., Féraud, R., and Bouneffouf, D. (2014). A neural networks committee for the contextual bandit problem. In *Proceedings of the 21st International Conference on Neural Information Processing (ICONIP)*, pages 374–381.

Audibert, J.-Y. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct) :2785–2836.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19) :1876–1902.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3) :235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1) :48–77.

Balmain, A., Ramsden, M., Bowden, G. T., and Smith, J. (1984). Activation of the mouse cellular harvey-ras gene in chemically induced benign skin papillomas. *Nature*, 307 :658–660.

Balzarotti, F., Eilers, Y., Gwosch, K. C., Gynnå, A. H., Westphal, V., Stefani, F. D., Elf, J., and Hell, S. W. (2016). Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science*, page aak9913.

Baransi, A., Maillard, O.-A., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In *Machine Learning and Knowledge Discovery in Databases*, pages 115–131.

Baumeister, R. F., Bratslavsky, E., Muraven, M., and Tice, D. M. (1998). Ego depletion : Is the active self a limited resource ? *Journal of Personality and Social Psychology*, 74(5) :1252.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8) :716–719.

Berning, S., Willig, K. I., Steffens, H., Dibaj, P., and Hell, S. W. (2012). Nanoscopy in a living mouse brain. *Science*, 335(6068) :551–551.

Bothe, M. K., Dickens, L., Reichel, K., Tellmann, A., Ellger, B., Westphal, M., and Faisal, A. A. (2013). The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Review of Medical Devices*, 10(5) :661–73.

Bowman Jr, V. J. (1976). On the relationship of the Tchebycheff norm and the efficient frontier of multiple-criteria objectives. In *Multiple criteria decision making*, pages 76–86.

Brochu, E., De Freitas, N., and Ghosh, A. (2008). Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 409–416.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1) :1–122.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, pages 23–37.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011a). X-armed bandits. *Journal of Machine Learning Research (JMLR)*, 12 :1655–1695.

Bubeck, S., Stoltz, G., and Yu, J. Y. (2011b). Lipschitz bandits without the Lipschitz constant. In *Algorithmic Learning Theory : 22nd International Conference (ALT)*, pages 144–158.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3) :1516–1541.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems 24 (NIPS)*, pages 2249–2257.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv :1706.03741*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 208–214.

Coello, C. A. C., Lamont, G. B., Van Veldhuizen, D. A., Goldberg, D. E., and Koza, J. R. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. 2nd edition.

Cotter, A., Keshet, J., and Srebro, N. (2011). Explicit approximations of the gaussian kernel. *arXiv preprint arXiv :1109.4603*.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17) :6889–6892.

Djolonga, J., Krause, A., and Cevher, V. (2013). High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1025–1033.

Drugan, M. M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms : A study. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Durand, A., Beaumont, J.-A., Gagné, C., Lemay, M., and Paquet, S. (2017a). Query completion using bandits for engines aggregation. In *3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A., Bordet, C., and Gagné, C. (2014). Improving the Pareto UCB1 algorithm on the multi-objective multi-armed bandit. In *Workshop on Bayesian Optimization at the 27th Neural Information Processing Systems (NIPS)*.

Durand, A. and Gagné, C. (2014). Thompson sampling for combinatorial bandits and its application to online feature selection. In *Proc. of the 28th AAAI Conference, Workshop on Sequential Decision-Making with Big Data*, pages 6–9.

Durand, A. and Gagné, C. (2017a). Estimating quality in user-guided multi-objective bandits optimization. *arXiv preprint arXiv :1701.01095*.

Durand, A. and Gagné, C. (2017b). Thompson sampling for user-guided multi-objective bandits optimization. In *3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A., Maillard, O.-A., and Pineau, J. (2017b). Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv :1708.00768*.

Durand, A. and Pineau, J. (2015a). Adaptive treatment allocation using sub-sampled gaussian processes. In *AAAI 2015 Fall Symposium on Embedded Machine Learning (EML)*.

Durand, A. and Pineau, J. (2015b). Cancer treatment optimization using gaussian processes. In *2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Durand, A. and Pineau, J. (2015c). Treatment allocation as contextual bandit. In *Workshop on Machine Learning in Healthcare at the 28th Neural Information Processing Systems (NIPS)*.

Ehrgott, M. (2012). Vilfredo pareto and multi-objective optimization. In *Documenta Mathematica – 21st International Symposium on Mathematical Programming*, pages 447–453.

Ernst, D., Stan, G., Goncalves, J., and Wehenkel, L. (2006). Clinical data based optimal sti strategies for hiv : a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672.

Escandell-Montero, P., Chermisi, M., Martínez-Martinez, J., Gomez-Sanchis, J., Barbieri, C., Soria-Olivas, E., Mari, F., Vila-Frances, J., Stopper, A., Gatti, E., and Martín-Guerrero, J. (2014). Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1) :47–60.

Féraud, R., Allesiardo, R., Urvoy, T., and Clérot, F. (2016). Random forest for the contextual bandit problem. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 93–101.

Fritzsche, M., Fernandes, R. A., Chang, V. T., Colin-York, H., Clausen, M. P., Felce, J. H., Galiani, S., Erlenkämper, C., Santos, A. M., and Heddleston, J. M. (2017). Cytoskeletal actin dynamics shape a ramifying actin network underpinning immunological synapse formation. *Science Advances*, 3(6) :e1603032.

Garivier, A.and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference on Learning Theory (COLT)*, pages 359–376.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. (2015). Bayesian reinforcement learning : A survey. *Foundations and Trends® in Machine Learning*, 8(5-6) :359–483.

Gittins, J. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266.

Göttfert, F., Pleiner, T., Heine, J., Westphal, V., Görlich, D., Sahl, S. J., and Hell, S. W. (2017). Strong signal increase in sted fluorescence microscopy by imaging regions of subdiffraction extent. *Proceedings of the National Academy of Sciences*, 114(9) :2125–2130.

Hanne, J., Göttfert, F., Schimer, J., Anders-Össwein, M., Konvalinka, J., Engelhardt, J., Müller, B., Hell, S. W., and Kräusslich, H.-G. (2016). Stimulated emission depletion nanoscopy reveals time-course of human immunodeficiency virus proteolytic maturation. *ACS nano*, 10(9) :8215–8222.

Hawkins, E. D. (2017). Advanced microscopy and imaging techniques in immunology and cell biology. *Immunology and Cell Biology*, 95(6) :499–500.

Hell, S. W. and Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission : stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11) :780–782.

Huang, B., Babcock, H., and Zhuang, X. (2010). Breaking the diffraction barrier : super-resolution imaging of cells. *Cell*, 143(7) :1047–1058.

Kan, A. (2017). Machine learning applications in cell image analysis. *Immunology and Cell Biology*, 95(6) :525–530.

Katariya, S., Kveton, B., Szepesvari, C., and Wen, Z. (2016). DCM bandits : Learning to rank with multiple clicks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1215–1224.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling : An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 199–213.

Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690.

Köksalan, M. M., Wallenius, J., and Zionts, S. (2011). *Multiple criteria decision making : from early history to the 21st century*. World Scientific.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2447–2455.

Krueger, D., Leike, J., Evans, O., and Salvatier, J. (2016). Active reinforcement learning : Observing rewards at a cost. In *Workshop on Future of Interactive Learning Machines at the 29th Neural Information Processing Systems (NIPS)*.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1) :4–22.

Laumanns, M., Thiele, L., Deb, K., and Zitzler, E. (2002). Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary computation*, 10(3) :263–82.

Li, C.-L., Kandasamy, K., Póczos, B., and Schneider, J. (2016). High dimensional bayesian optimization via restricted projection pursuit models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 884–892.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web (WWW)*, pages 661–0670.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning in genetics and genomics. *Nature Reviews Genetics*, 16(6) :321–332.

Loizides, C., Iacovides, D., Hadjiandreou, M. M., Rizki, G., Achilleos, A., Strati, K., and Mitsis, G. D. (2015). Model-based tumor growth dynamics and therapy response in a mouse model of de novo carcinogenesis. *PloS One*, 10(12) :e0143840.

Maghsudi, S. and Hossain, E. (2016). Multi-armed bandits with application to 5G small cells. *IEEE Wireless Communications*, 23(3) :64–73.

Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits : Regret lower bound and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 975–999.

Maillard, O.-A. (2016). Self-normalization techniques for streaming confident regression. https://hal.archives-ouvertes.fr/hal-01349727.

Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th annual Conference on Learning Theory (COLT)*, pages 497–514.

Marchant, R. and Ramos, F. (2014). Bayesian optimisation for informative continuous path planning. In *International Conference on Robotics and Automation (ICRA)*, pages 6136–6143. IEEE.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*.

Miettinen, K. (2014). Survey of methods to visualize alternatives in multiple criteria decision making problems. *OR spectrum*, pages 1–35.

Miller, H. C., Pattison, K. F., DeWall, C. N., Rayburn-Reeves, R., and Zentall, T. R. (2010). Self-control without a "self"? common self-control processes in humans and dogs. *Psychological science*, 21(4) :534–538.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1) :62–66.

Panuccio, G., Guez, A., Vincent, R., Avoli, M., and Pineau, J. (2013). Adaptive control of epileptiform excitability in an in vivo model of limbic seizures. *Experimental Neurology*, 241 :179–83.

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2) :693–721.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT press Cambridge.

Rigollet, P. (2015). 18.S997 High-Dimensional Statistics, Chapter 1, Spring 2015. (MIT OpenCourseWare : Massachusetts Institute of Technology), https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015/lecture-notes/MIT18_S997S15_Chapter1.pdf. (Accessed March 15, 2017). License : Creative commons BY-NC-SA.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 54–66.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5) :527–535.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4) :1221–1243.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research (JMLR)*, 17(68) :1–30.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6) :639–658.

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research (JMLR)*, 15 :2533–2568.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 2951–2959.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M. A., Prabhat, and Adams, R. (2015). Scalable bayesian optimization using deep neural networks. In *Proceedings of the 32st International Conference on Machine Learning (ICML)*, pages 2171–2180.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 4134–4142.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting : No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5) :3250–3265.

Staudt, T., Engler, A., Rittweger, E., Harke, B., Engelhardt, J., and Hell, S. W. (2011). Far-field optical nanoscopy with reduced number of state transition cycles. *Optics express*, 19(6) :5644–5657.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4) :285–294.

Tomayko, M. M. and Reynolds, C. P. (1989). Determination of subcutaneous tumor size in athymic (nude) mice. *Cancer chemotherapy and pharmacology*, 24(3) :148–154.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th conference on Uncertainty In Artificial Intelligence (UAI)*, pages 654–665.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials : benefits and challenges. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 30(2) :199.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., and Tice, D. M. (2008). Making choices impairs subsequent self-control : A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology*, 94(5) :883–898.

Wang, Z. and de Freitas, N. (2014). Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv :1406.7758*.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Feitas, N. (2016). Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55 :361–387.

Williamson, S. F., Jacko, P., Villar, S. S., and Jaki, T. (2017). A bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics & Data Analysis*, 113 :136–153.

Willig, K. I., Kellner, R., Medda, R., Hein, B., Jakobs, S., and Hell, S. W. (2006). Nanoscale resolution in gfp-based microscopy. *Nature Methods*, 3(9) :721–723.

Wilson, A., Fern, A., and Tadepalli, P. (2014). Using trajectory data to improve bayesian optimization for reinforcement learning. *Journal of Machine Learning Research*, 15 :253–282.

Winter, F. R., Loidolt, M., Westphal, V., Butkevich, A. N., Gregor, C., Sahl, S. J., and Hell, S. W. (2017). Multicolour nanoscopy of fixed and living cells with a single sted beam and hyperspectral detection. *Scientific Reports*, 7.

Yahyaa, S. Q., Drugan, M. M., and Manderick, B. (2015). Thompson sampling in the adaptive linear scalarized multi objective multi armed bandit. In *Proceedings of the 7th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 55–65.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2009). The $k$-armed dueling bandits problem.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2012). The $k$-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5) :1538–1556.

Zhao, Y., Kosorok, M., and D., Z. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26) :3294–315.

Zuluaga, M., Sergent, G., Krause, A., and Püschel, M. (2013). Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 462–470.